Evaluating Metric Sensitivity to Offline–Online Alignment in Information Retrieval

Satsuki Udagawa
s.udagawa@student.tudelft.nl
Responsible Professor: Avishek Anand

EEMCS, Delft University of Technology, The Netherlands

Background and Motivation

Common Methods to Evaluate Information Retrieval(IR) Systems:

- **Offline Metrics (e.g., MAP, nDCG):** Fast and reproducible, but rely on static relevance labels that may not reflect real user behaviour.
- **Online Metrics (e.g., CTR, dwell time):** Capture genuine user interactions, but are costly, slow, and often unavailable during development.

The Alignment Problem: Offline improvements do not always translate into better online performance, creating uncertainty in model selection.

Missing Piece: Sensitivity Most prior work studies *correlation*, but not *how strongly* offline metrics respond when online behaviour changes.

Research Question: How sensitive are common offline IR metrics to changes in offline–online alignment?

Methodology

Data & Systems:

- 52 diverse ranking pipelines (lexical, dense, multi-vector, sparse).
- MS MARCO DL19 & DL20 queries with high-quality relevance labels.

Offline Metrics:

- Precision@10, Recall@10, MAP, MRR, nDCG@10.

Online Metrics (Simulated User Behaviour):

- Click through rate (CTR), Session success rate (SSR), Zero result rate (ZRR), Average dwell time (ADT), Session abandonment rate (SAR).

Analysis Procedure:

- Compute all offline and online metrics for each system.
- For every offline–online pair, plot scores and fit a least-squares line.
- **Sensitivity:** Absolute slope of the fitted line.
- **Alignment:** Pearson correlation across systems.

Results

Sensitivity (Slopes):

	CTR	SSR	ZRR	ADT	SAR
MAP	7.40	0.525	-0.525	-0.00109	0.117
MRR	7.48	0.529	-0.529	-0.00110	0.118
NDCG@10	8.16	0.595	-0.595	-0.00116	0.107
Precision@10	0.813	0.0886	-0.0886	-0.000108	0.00149
Recall@10	7.76	0.845	-0.845	-0.00103	0.0130

Sensitivity of offline metrics (slopes).

Alignment (Correlation):

	CTR	SSR	ZRR	ADT	SAR
MAP	0.521	0.846	-0.846	-0.296	0.169
MRR	0.523	0.845	-0.845	-0.297	0.168
NDCG@10	0.523	0.872	-0.872	-0.286	0.141
Precision@10	0.385	0.958	-0.958	-0.197	0.014
Recall@10	0.385	0.957	-0.957	-0.196	0.013

Linear alignment (Pearson correlation).

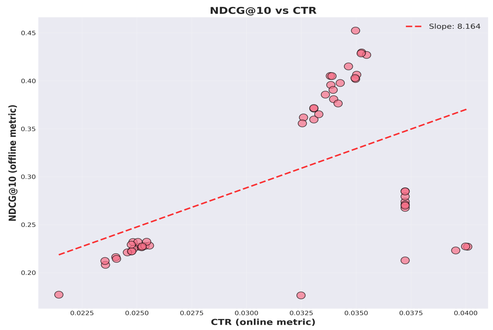

Figure 1: NDCG@10 vs CTR scatter plot.

The scatter plot in **Figure 1** illustrates how the offline metric NDCG@10 varies with the online metric CTR across all 52 ranking systems. Each point represents one system, with CTR on the x-axis and NDCG@10 on the y-axis. The red line shows the least-squares fit, whose positive slope indicates that systems with higher CTR tend to also achieve higher NDCG@10 scores. The spread of points around the line reflects how consistently (or inconsistently) NDCG@10 tracks changes in CTR: a tight cluster would indicate strong alignment, while wider dispersion suggests that the relationship is present but not perfectly linear.

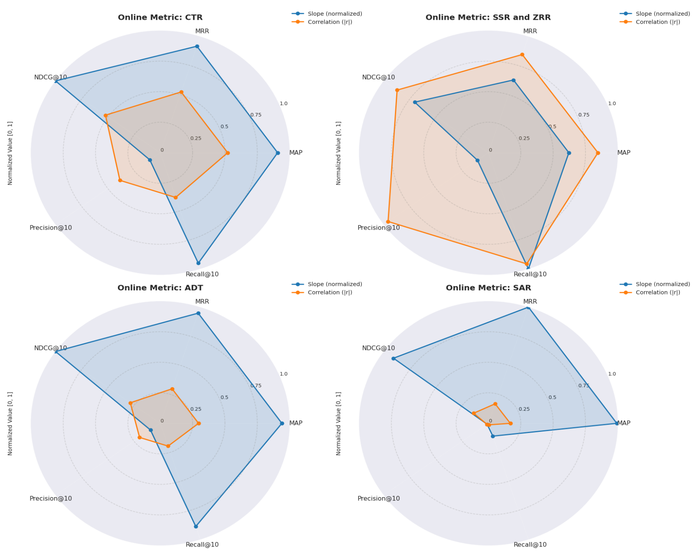

Figure 2: Radar charts of normalized sensitivity and absolute correlations.

How to interpret these results: Because the online metrics operate on different numerical scales, comparisons of sensitivity (slopes) should be made **within each online metric column**, not across columns.

In the radar charts, slope values are **normalized per online metric** so that the offline metric with the steepest slope receives a value of 1. Pearson correlations are shown as **absolute values** to avoid negative values in the radar visualization.

Conclusions

- **Offline metrics differ substantially in sensitivity.** NDCG@10 and Recall@10 respond most strongly to changes in online behaviour across multiple signals.
- **Precision@10 is consistently insensitive.** Its slopes are an order of magnitude smaller than other metrics, making it unreliable for detecting meaningful performance differences.
- **Sensitivity and alignment capture different properties.** Some metric pairs show steep slopes but weak correlations, meaning strong reactions without consistent tracking of online performance.
- **No single offline metric is universally best.** Different online behaviours emphasize different aspects of ranking quality, meaning that metric choice should reflect the target user signal.

Limitations & Future Work

- **Non-linear offline–online relationships.** Several offline– online metric pairs do not follow a clear linear trend; further work is needed to understand the underlying relationships.
- **Simulated online behaviour.** Online metrics are derived from simulation models, which may not fully capture real user behaviour or noise patterns.
- **Limited behavioural diversity.** The five online metrics cover common signals, but do not cover other variables such as multi-click sessions, user satisfaction, or effects of long-term engagement.
- **Metric coverage.** Only five offline metrics were evaluated; extending to ERR, RBP, or alternative cutoffs (different values for k) could test generality.