How Well do Offline Metrics Predict Online Performance of Product Ranking Models?

Xiaojie Wang*
xiojie@amazon.com
Amazon.com
Palo Alto, CA, United States

Ruoyuan Gao*
ruoyuag@amazon.com
Amazon.com
Palo Alto, CA, United States

Anoop Jain
anoopjai@amazon.com
Amazon.com
Palo Alto, CA, United States

Graham Edge
edgegrah@amazon.com
Amazon.com
Seattle, WA, United States

Sachin Ahuja
satchmo@amazon.com
Amazon.com
Palo Alto, CA, United States

ABSTRACT

Online evaluation techniques are widely adopted by industrial search engines to determine which ranking models perform better under a certain business metric. However, online evaluation can only evaluate a small number of rankers and people resort to offline evaluation to select rankers that are likely to yield good online performance. To use offline metrics for effective model selection, a major challenge is to understand how well offline metrics predict which ranking models perform better in online experiments. This paper aims to address this challenge in product search ranking. Towards this end, we collect gold data in the form of preferences over ranker pairs under a business metric in e-commerce search engine. For the first time, we use such gold data to evaluate offline metrics in terms of directional agreement with the business metric. Furthermore, we analyze offline metrics in terms of discriminative power through paired sample t-test and rank correlations among offline metrics. Through extensive online and offline experiments, we studied 36 offline metrics and observed that: (1) Offline metrics align well with online metrics: they agree on which one of two ranking models is better up to 97% of times; (2) Offline metrics are highly discriminative on large-scale search ranking data, especially NDCG (Normalized Discounted Cumulative Gain) which has a discriminative power over 99%.

CCS CONCEPTS

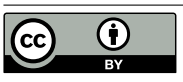
• **Information systems → Retrieval effectiveness**.

KEYWORDS

Evaluation metrics; online evaluation; offline evaluation

ACM Reference Format:
Xiaojie Wang, Ruoyuan Gao, Anoop Jain, Graham Edge, and Sachin Ahuja. 2023. How Well do Offline Metrics Predict Online Performance of Product Ranking Models? . In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3539618.3591865

*Both authors contributed equally to this research.



SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591865

1 INTRODUCTION

A common problem in evolving a search engine in industry is deciding which ranking models perform better under certain business metrics [7, 9, 24]. The best ranking models are continuously deployed to the search engine in production to serve users' queries in real-world. The most reliable way to assess model performance is arguably to compare models online to collect real user feedback. Popular online evaluation techniques include A/B testing [20, 23], and interleaving methods [12] such as Balanced interleaving [6, 22] and TeamDraft interleaving [34]. Given two ranking models A and B, A/B testing randomly assigns each user to either ranker A (control) or ranker B (treatment), and calculates business metrics for each model based on collected user feedback (e.g., clicks and purchases). Interleaving differs from A/B in that it measures the relative performance difference between A and B, where a single merged ranking result of A and B is presented to the same user. Through statistical analysis on the difference in online performance between A and B, we decide that A is better or worse than B with a certain p-value under each business metric, e.g., click-through rate [27]. Although online evaluation of ranking models is reliable, it has some limitations: (1) We can only compare a limited number of models within a time period because each online experiment requires a large amount of user feedback to reach statistically significant conclusions while user feedback is scarce resource; (2) Online experiment has the risk of deteriorating user search experiences and losing user trusts if testing a potentially poor ranking model.

Due to the above limitations with online evaluation, offline evaluation is widely adopted to select candidate models that are worth testing online [10, 30, 31, 49]. Different from online evaluation, the number of models that can be compared by offline evaluation is not limited by the volume of real-time user feedback. Offline evaluation usually uses a set of queries, a set of products for each query, relevance labels for the products with respect to the query, and some offline metrics. Each offline metric takes a ranked list of products produced by a ranking model, and computes a numeric score based on relevance labels [8, 29, 32, 37]. A major challenge in offline evaluation is to identify which offline metrics agree the most with online business metrics from the viewpoint of which ranking model is better. Using offline metrics with a higher agreement with

online metrics, models selected offline are more likely to yield good results when tested online. Note that we will use online metrics and business metrics interchangeably in this paper.

How well offline metrics agree with business metrics in preferences of ranking models has not been well-studied in literature. A closely related line of research is to study how well offline metrics agree with human judges' satisfactions with search engine result pages (SERPs) [4, 39, 40]. E.g., some works study whether the numeric score that an offline metric assigns to an SERP is a good predictor of human judges' satisfactions [3, 26, 28], while other works study whether an offline metric agrees with human judges in preference of an SERP over another [38]. In these works, the gold data used to study offline metrics contains either scores of satisfactions with an SERP or preferences of an SERP over another. However, such gold data does not have preferences of a ranking model over another, which are what we collect and use to study offline metrics in this paper. Besides, existing works usually hire human judges to obtain the gold data in a laboratory setting, which has the following limitations: (1) The gold data is small (hundreds or thousands of SERPs) mostly because human judging is costly; (2) It is subject to selection bias in the hiring process and cognition bias of hired human judges. In comparison, we collect gold data by aggregating real users' feedback on billions of SERPs through multiple interleaving experiments in a production setting.

In this paper, we study directional agreement between offline and business metrics in terms of preferences over ranking models for product search. To obtain gold data for evaluating offline metrics, we compare online performance of 114 ranker pairs through Balanced interleaving. We measure the online performance by aggregating real users' actions over billions of queries under a business metric called Purchased Units [50]. For the first time, we use such gold data to measure how well offline metrics predict which one of two rankers has better online performance. For completeness, we also measure the stability of offline metrics across search queries and experiments using discriminative power based on paired t-test. Moreover, we adopt Kendall's τ [25] and symmetric τ_{ap} [48] to study rank correlation between offline metrics. We study 36 offline metrics through extensive experiments at Amazon Search. Our main results are that: (1) Offline metrics align well with Purchased Units: they agree on which one of two ranking models is better up to 97% of times; (2) Offline metrics are highly discriminative on a large-scale dataset, especially NDCG (Normalized Discounted Cumulative Gain) which has a discriminative power over 99%.

2 METRICS CONSIDERED IN OUR STUDY

First, we describe an online metric Purchased Units in Section 2.1. Then, we describe two widely used classes of offline metrics, Expected Utility and Normalized Discounted Cumulative Gain, in Section 2.2. We derive 36 instances of the offline metrics by using different relevance labels, rank position cutoffs, gain functions, etc.

2.1 Online Metrics

Online performance of a ranking model is measured on search logs consisting of search results and user feedback. Suppose the search logs consist of n search queries issued by users, $\mathcal{L} = \{q_1, q_2, ..., q_n\}$. For each query q_i, the model returns a ranking list of m_i products $\{p_{i,1}, p_{i,2}, ..., p_{i,m_i}\}$. In this paper, we measure the model's online performance by Purchased Units [50], defined by the total number of products purchased within a certain time period:

$$\text{Purchased Units} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \texttt{qty}(q_i, p_{i,j}). \tag{1}$$

Here, $\texttt{qty}(q, p) \in [0, 1, 2, ...]$ is the quantity of product p purchased by a user given query q. There are a wide range of other online metrics [16, 44], which we leave for future work.

2.2 Offline Metrics

For offline evaluation, we use an offline dataset created from historical search logs. With slight abuse of notation, we use n to denote the number of queries in the offline dataset, $\mathcal{D} = \{q_1, q_2, ..., q_n\}$, and use $p_{i,j}$ to denote the j-th product in the ranking list of query q_i induced by the ranking model.

Expected Utility (EU): As Purchased Units does not consider the positions of purchased products, we first examine EU [5], which assigns equal importance to each of the top-k positions:

$$\text{EU@k} = \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{k} y_{i,j}}{k} \right). \tag{2}$$

Here, $y_{i,j}$ is a relevance label of product $p_{i,j}$ for query q_i.

Normalized Discounted Cumulative Gain (NDCG): Products ranked at higher positions usually receive more user attention and have higher purchase probabilities [18, 19, 47]. To account for such position bias, we consider NDCG [21] which rewards relevant products being ranked at higher positions. First, we define Discounted Cumulative Gain (DCG) for a ranking list of query q_i:

$$\text{DCG}(k, q_i) = \sum_{j=1}^{k} \frac{\texttt{gain}(y_{i,j})}{\log_2(1 + j)}. \tag{3}$$

Here, $\texttt{gain}(\cdot)$ is a gain function of linear form $\texttt{gain}(y) = y$ or exponential form $\texttt{gain}(y) = 2^y - 1$. Then, we normalize DCG to $[0, 1]$ for each query and define NDCG by summing the normalized DCG values for all queries:

$$\text{NDCG}(k) = \sum_{i=1}^{n} \frac{\text{DCG}(k, q_i)}{\text{maxDCG}(k, q_i)}, \tag{4}$$

where maxDCG is the DCG value of the ranking list obtained by sorting products in descending order of relevance.

We explore three methods of defining relevance labels $y_{i,j}$: (1) Binary purchase $\texttt{binary}(q_i, p_{i,j}) \in \{0, 1\}$ indicates whether product $p_{i,j}$ is purchased for query q_i [15]; (2) Purchase quantity $\texttt{qty}(q_i, p_{i,j}) \in [0, \infty]$ means how many units of product $p_{i,j}$ are purchased for query q_i [41]; (3) Purchase probability $\texttt{prob}(q_i, p_{i,j}) \in [0, 1]$ means the probability of product $p_{i,j}$ being purchased for same query q_i. We compute purchase probabilities by dividing the number of purchases of a product by the number of times it was shown to users [9] and use a position-based model (PBM) [1, 13] trained on expectation–maximization (EM) algorithm [2, 43] to debias the purchase probabilities. To denoise relevance labels, we can discretize their values into integers $[0, 1, ..., b]$ given number of buckets b: we divide the values by the maximum value for each query, multiply by b, and round the resulting values to the closest integers.

Table 1: Agreement of offline metrics w.r.t. online metric Purchased Units.

	α with 95% CI for all 114 ranker pairs		α with 95% CI for 101 sig. diff. ranker pairs	
	$k = 10$	$k = 20$	$k = 10$	$k = 20$
EU(binary)	0.921 [0.857, 0.958]	0.895 [0.825, 0.939]	**0.970 [0.916, 0.990]**	0.941 [0.876, 0.972]
EU(qty, $b \in \{10, 20\}$)	0.921 [0.857, 0.958]	0.904 [0.835, 0.945]	**0.970 [0.916, 0.990]**	0.950 [0.889, 0.979]
EU(qty, $b = \infty$)	0.930 [0.868, 0.964]	0.895 [0.825, 0.939]	**0.970 [0.916, 0.990]**	0.941 [0.876, 0.972]
EU(prob, $b \in \{10, 20\}$)	0.728 [0.640, 0.801]	0.675 [0.585, 0.754]	0.772 [0.681, 0.843]	0.733 [0.639, 0.809]
EU(prob, $b = \infty$)	0.711 [0.621, 0.786]	0.623 [0.531, 0.706]	0.743 [0.650, 0.818]	0.663 [0.567, 0.748]
NDCG(binary)	**0.939 [0.879, 0.970]**	0.930 [0.868, 0.964]	**0.970 [0.916, 0.990]**	**0.970 [0.916, 0.990]**
NDCG(lin, qty, $b \in \{10, 20, \infty\}$)	**0.939 [0.879, 0.970]**	0.930 [0.868, 0.964]	**0.970 [0.916, 0.990]**	**0.970 [0.916, 0.990]**
NDCG(lin, prob, $b \in \{10, 20, \infty\}$)	0.728 [0.640, 0.801]	0.711 [0.621, 0.786]	0.772 [0.681, 0.843]	0.752 [0.660, 0.826]
NDCG(exp, qty, $b \in \{10, 20\}$)	**0.939 [0.879, 0.970]**	0.930 [0.868, 0.964]	**0.970 [0.916, 0.990]**	**0.970 [0.916, 0.990]**
NDCG(exp, prob, $b = 10$)	0.737 [0.649, 0.809]	0.746 [0.659, 0.817]	0.772 [0.681, 0.843]	0.782 [0.692, 0.852]
NDCG(exp, prob, $b = 20$)	0.754 [0.668, 0.824]	0.754 [0.668, 0.824]	0.782 [0.692, 0.852]	0.782 [0.692, 0.852]

3 EVALUATION OF OFFLINE METRICS

In this section, we describe the techniques that we use to evaluate offline metrics. Section 3.1 introduces online agreement to measure how well results of offline metrics agree with those of business metrics. Section 3.2 describes two types of techniques that are widely used to evaluate offline metrics given an offline dataset.

3.1 Online Agreement of Offline Metrics

The most important measurement that we use to evaluate offline metrics is called online consistency.

Data collection of online ground truth. It is challenging to collect sufficient online data based on real user feedback as the gold data of preferences among ranking models. With traditional A/B testing, this is time-consuming and puts user experience at risk if the tested models are bad. To tackle this challenge, we utilize interleaving experiment which is orders of magnitude more sensitive than traditional A/B tests [6]. This way, we can collect more online results with less customer traffic and shorter experiment duration. This also enables the feasibility to collect most recent data to build accurate offline to online mapping. Through this method, we obtain online user preference data. Each data point is a relative comparison between a pair of rankers on a given online metric. We study offline and online agreement by looking at all ranker pairs as well as only pairs that have statistically significant online difference.

Definition of offline-online agreement measure. Because online preference aims to identify the better ranker among a pair of rankers, we focus on directional agreement between offline and online evaluation. Although our online data is collected through interleaving, the same method can also be applied to A/B testing results where we can obtain pairwise preferences by comparing the absolute value of online metrics. Agreement reflects how likely an offline metric can correctly predict the online preference. Formally, let N denote the total number of ranker pairs evaluated online. Given an offline metric and an online metric, let C denote the number of concordant pairs that, orderings of the same ranker pair are the same by the offline metric and by the online metric. We compute agreement α as the proportion of concordant pairs: $\alpha = C/N$. To compute confidence intervals, we assume that agreement results follow a Binomial distribution, where agreement with online Purchased Units on each ranker pair is a binary outcome. As we have a relatively small number of comparison results, using Normal approximation interval may become inappropriate and can cause overshoot problem [33] where the confidence intervals exceed the range of $[0, 1]$. Instead, we adopt Wilson score interval [46]. To account for selection bias in ranker pairs, we apply bootstrapping and obtain the confidence intervals with respect to different number of resamples. An alternative to the agreement measure is Goodman and Kruskal's γ correlation [14]. It can be shown that the α and γ are equivalent in measuring offline-online agreement as $\gamma = 2\alpha - 1$.

3.2 Offline Evaluation of Offline Metrics

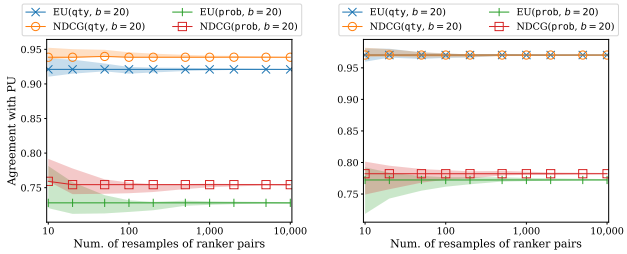
Discriminative power measures the stability of an offline metric across queries and experiments using significance test with a given significance level [36]. For an offline metric, being highly discriminative is a preferred characteristic because it means that the metric can identify subtle differences between ranking models. Existing works use paired bootstrap test [35] or Tukey's Honestly Significant Differences (HSD) test [11] to calculate discriminative power on small testing data that contains hundreds of queries. Because our testing data contains millions of queries which is relatively large-scale, we choose paired sample t-test. For each offline metric, we compute the percentage of statistically significant ranker pairs with respect to p-value < 0.05 and the observed minimum delta between two rankers that are found to be statistically significant [38].

We measure correlation between two offline metrics by using each metric to rank models from best to worse, and then calculating rank correlation between the resulting two ranking lists of models. Rank correlation is useful to study which metrics are the most different or similar in terms of deciding which models perform better [17, 45]. A widely used statistic, Kendall's τ, is symmetric but lacks a property called top heaviness, meaning that it treats swaps near top of a ranking the same as those near bottom. However, in search ranking, swaps near top of a ranking are generally more important than those near bottom. Yilmaz et al. propose τ_{ap} [48] to account for top heaviness. However, τ_{ap} is not symmetric and we adopt symmetric τ_{ap} [42] in our study. Both Kendall's τ and symmetric τ_{ap} range from -1 to 1, where -1 means that two rankings always disagree and 1 means that two rankings are identical.

4 EXPERIMENTS

4.1 Datasets

We obtain online performance of 114 ranker pairs in terms of Purchased Units via multiple interleaving experiments. There are 113 distinct ranking models in total. Our gold data is composed



(a) All ranker pairs (b) Sig. diff. ranker pairs

Figure 1: Bootstrap confidence interval of online agreement.

of relative Purchased Units differences for the 114 ranker pairs, among which 101 are statistically significantly different with p-value < 0.05.

We collect offline testing data by replaying actual users' queries and, to achieve a high fidelity of the data, we replay queries shortly after the actual queries were issued. The testing data consists of around two million queries. We compute a total of 36 offline metrics by using different: (1) relevance labels: binary purchase (`binary`), purchase quantity (`qty`), or purchase probability (`prob`); (2) gain functions: linear gain (`lin`) or exponential gain (`exp`); (3) number of buckets for discretizing labels: $b \in \{10, 20, \infty\}$ where $b = \infty$ means no discretization; (4) rank position cutoffs $k \in \{10, 20\}$. Note that we do not compute NDCG with exponential gain without discretization to avoid numeric overflow.

4.2 Evaluation Results

Agreement with online performance. Table 1 shows the online agreement of offline metrics and Wilson score interval at 95% confidence level. In our case, the best offline metrics achieve online agreement up to 0.970. In other words, we can correctly predict online preferences of ranker pairs up to 97% of the time. Figure 1 shows bootstrap confidence intervals of online agreement against the number of resamples. We omit bootstrap results of metrics using binary purchase as the results are consistent with those using purchase quantity. We can see that the upper bound of metrics using purchase probability is lower than that of metrics using binary purchase or purchase quantity labels. Besides, NDCG metrics have higher online agreement than EU metrics, and metrics using rank position cutoff $k = 10$ have higher online agreement than those using $k = 20$. These findings highlight the importance of considering position bias at higher rank positions. We can also see that metrics using discretization generally better agree with online evaluation because discretization helps reduce the noise in relevance labels.

Discriminative power. Table 2 shows discriminative power for all 36 offline metrics. We conduct paired t-test for all pairwise comparisons between the 113 rankers (113 * 112 / 2 = 6328 pairs). We observe that: (1) Discriminative power of the offline metrics is generally high (the lowest is 96.4%); (2) NDCG metrics generally have a higher discriminative power than EU metrics as expected; (3) Metrics evaluated at $k = 10$ have a larger discriminative power than those evaluated at $k = 20$ metrics; (4) Using probability of purchase is more discriminative than using binary purchase or purchase quantity due to a larger cardinality.

Table 2: Discriminative power at 0.05 significance level.

	$k = 10$		$k = 20$	
	Discr. pwr.	Min Δ	Discr. pwr.	Min Δ
EU(`binary`)	98.1%	0.000029	96.9%	0.000006
EU(`qty`, $b = 10$)	98.1%	0.000296	96.9%	0.000057
EU(`qty`, $b = 20$)	98.1%	0.000591	96.9%	0.000114
EU(`qty`, $b = \infty$)	97.9%	0.000033	96.4%	0.000009
EU(`prob`, $b = 10$)	99.2%	0.000447	98.8%	0.000109
EU(`prob`, $b = 20$)	99.2%	0.000606	98.8%	0.000299
EU(`prob`, $b = \infty$)	99.1%	0.000015	97.9%	0.000006
NDCG(`binary`)	99.0%	0.000131	98.9%	0.000160
NDCG(`lin`, `qty`, $b = 10$)	99.0%	0.000132	98.9%	0.000160
NDCG(`lin`, `qty`, $b = 20$)	99.0%	0.000132	98.9%	0.000160
NDCG(`lin`, `qty`, $b = \infty$)	99.0%	0.000132	98.9%	0.000160
NDCG(`lin`, `prob`, $b = 10$)	99.3%	0.000089	99.2%	0.000091
NDCG(`lin`, `prob`, $b = 20$)	**99.4%**	0.000087	99.2%	0.000085
NDCG(`lin`, `prob`, $b = \infty$)	99.3%	0.000088	99.3%	0.000081
NDCG(`exp`, `qty`, $b = 10$)	99.0%	0.000133	98.9%	0.000160
NDCG(`exp`, `qty`, $b = 20$)	99.0%	0.000133	98.9%	0.000159
NDCG(`exp`, `prob`, $b = 10$)	99.2%	0.000139	**99.4%**	0.000099
NDCG(`exp`, `prob`, $b = 20$)	99.1%	0.000136	99.1%	0.000119

Table 3: Kendall's τ / Symmetric τ_{ap} between NDCG metrics.

	NDCG (`lin`, `qty`)	NDCG (`lin`, `prob`)	NDCG (`exp`, `qty`)	NDCG (`exp`, `prob`)
NDCG(`binary`)	1.000/1.000	0.364/0.260	1.000/1.000	0.485/0.409
NDCG(`lin`, `qty`)	-	0.364/0.260	1.000/1.000	0.485/0.409
NDCG(`lin`, `prob`)	-	-	0.364/0.260	0.819/0.741
NDCG(`exp`, `qty`)	-	-	-	0.485/0.408

Rank correlation. Table 3 shows rank correlations between the NDCG@10 metrics that have highest online agreement and discriminative power ($b = 20$ and $k = 10$). We observe that (1) two NDCG metrics using purchase probability are least correlated with the other three NDCG metrics using binary purchase or purchase quantity, e.g., Kendall's τ between NDCG(`lin`, `prob`) and NDCG(`binary`) is 0.364; (2) NDCG metrics using binary purchase and quantity purchase have perfectly positive correlations.

5 CONCLUSIONS

In this paper, we studied a total of 36 offline metrics for product search ranking. First, we introduced our methodologies to evaluate how well offline metrics predict online performance of ranking models. Specifically, we described how to collect online gold data and use the gold data for computing online agreement, which has challenges particular to an industry setting. Then, we studied discriminative power of offline metrics and rank correlation between offline metrics on large-scale search ranking data. Through extensive experiments, we observed that: (1) Offline metrics have high agreement with online Purchased Units: they agree on which one of two ranking models is better up to 97% of times; (2) Offline metrics are highly discriminative on large-scale search ranking data, especially NDCG whose discriminative power is over 99.0%.

ACKNOWLEDGMENTS

To Daria Sorokina and Micah Hodosh for suggestions that helped improve the quality of this work. To Meixue Yuan, Luochao Wang, and the rest of the team for their support in building the infrastructure and tools that made this work possible.

COMPANY PORTRAIT

Amazon.com is an e-commerce company that sells a wide selection of products, including electronics, groceries, etc, to customers worldwide. Amazon e-commerce websites are country-specific, e.g., amazon.com for the U.S. and amazon.co.uk for the U.K.. Amazon Search team develops and operates product search engines that serve product search queries on all Amazon e-commerce websites all over the world.

PRESENTER BIO

Xiaojie Wang is an Applied Scientist at Amazon in Amazon Search team. He received B.S. degree from the Renmin University of China in 2016 and Ph.D. degree from University of Melbourne in 2020. His research interests include information retrieval, recommmender systems, and machine learning. He has published more than 10 papers in prestigious conferences and journals including SIGIR, WSDM, NeurIPS, ICML, TKDE, and TPAMI. He has been regularly serving as a reviewer in top conferences and journals including CIKM, NeurIPS, ICML, and JMLR.

REFERENCES

[1] Aman Agarwal, Xuanhui Wang, Cheng Li, Michael Bendersky, and Marc Najork. 2019. Addressing trust bias for unbiased learning-to-rank. In *The World Wide Web Conference*. 4–14.

[2] Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W Bruce Croft. 2018. Unbiased learning to rank with unbiased propensity estimation. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 385–394.

[3] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 773–774.

[4] Azzah Al-Maskari, Mark Sanderson, Paul Clough, and Eija Airio. 2008. The good and the bad system: does the test collection predict users' effectiveness?. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 59–66.

[5] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the utility of search engine result pages: an information foraging based measure. In *The 41st International ACM SIGIR conference on research & development in information retrieval*. 605–614.

[6] Nan Bi, Pablo Castells, Daniel Gilbert, Slava Galperin, Patrick Tardif, and Sachin Ahuja. 2022. Debiased balanced interleaving at Amazon Search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2913–2922.

[7] Eliot P Brenner, Jun Zhao, Aliasgar Kutiyanawala, and Z Yan. 2018. End-to-end neural ranking for ecommerce product search. *Proceedings of SIGIR eCom* 18 (2018).

[8] C E Buckley and Ellen M Voorhees. 2005. Retrieval system evaluation. (2005).

[9] David Carmel, Elad Haramaty, Arnon Lazerson, and Liane Lewin-Eytan. 2020. Multi-objective ranking optimization for product search using stochastic label aggregation. In *Proceedings of The Web Conference 2020*. 373–383.

[10] Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval*. 903–912.

[11] Benjamin A Carterette. 2012. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Transactions on Information Systems (TOIS)* 30, 1 (2012), 1–34.

[12] Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. 2012. Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems (TOIS)* 30, 1 (2012), 1–41.

[13] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. Click models for web search. *Synthesis lectures on information concepts, retrieval, and services* 7, 3 (2015), 1–115.

[14] Herbert L Costner. 1965. Criteria for measures of association. *American Sociological Review* (1965), 341–353.

[15] Yaron Fairstein, Elad Haramaty, Arnon Lazerson, and Liane Lewin-Eytan. 2022. External Evaluation of Ranking Models under Extreme Position-Bias. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 252–261.

[16] Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. 2022. KuaiRand: An Unbiased Sequential Recommendation Dataset with Randomly Exposed Videos. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3953–3957.

[17] Ruoyuan Gao, Yingqiang Ge, and Chirag Shah. 2022. FAIR: Fairness-aware information retrieval evaluation. *Journal of the Association for Information Science and Technology* 73, 10 (2022), 1461–1473.

[18] Ruocheng Guo, Xiaoting Zhao, Adam Henderson, Liangjie Hong, and Huan Liu. 2020. Debiasing grid-based product search in e-commerce. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2852–2860.

[19] Siyuan Guo, Lixin Zou, Yiding Liu, Wenwen Ye, Suqi Cheng, Shuaiqiang Wang, Hechang Chen, Dawei Yin, and Yi Chang. 2021. Enhanced doubly robust learning for debiasing post-click conversion rate estimation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 275–284.

[20] Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. 2019. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter* 21, 1 (2019), 20–35.

[21] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* (2002), 422–446.

[22] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 133–142.

[23] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2017. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1517–1525.

[24] Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. 2017. On application of learning to rank for e-commerce search. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 475–484.

[25] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.

[26] Rohan Kumar, Mohit Kumar, Neil Shah, and Christos Faloutsos. 2018. Did we get it right? Predicting query performance in e-commerce search. *arXiv preprint arXiv:1808.00239* (2018).

[27] Pan Li. 2021. Leveraging Multi-Faceted User Preferences for Improving Click-Through Rate Predictions. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 864–868.

[28] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When does relevance mean usefulness and user satisfaction in web search?. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 463–472.

[29] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 1–38.

[30] Alistair Moffat, Joel Mackenzie, Paul Thomas, and Leif Azzopardi. 2022. A flexible framework for offline effectiveness metrics. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 578–587.

[31] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 659–668.

[32] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 1–27.

[33] Robert G Newcombe. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine* 17, 8 (1998), 857–872.

[34] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How does clickthrough data reflect retrieval quality?. In *Proceedings of the 17th ACM conference on Information and knowledge management*. 43–52.

[35] Tetsuya Sakai. 2006. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 525–532.

[36] Tetsuya Sakai. 2012. Evaluation with informational and navigational intents. In *Proceedings of the 21st international conference on World Wide Web*. 499–508.

[37] Tetsuya Sakai and Zhicheng Dou. 2013. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 473–482.

[38] Tetsuya Sakai and Zhaohao Zeng. 2019. Which Diversity Evaluation Measures Are" Good"?. In *Proceedings of the 42nd international ACM SIGIR conference on Research and Development in information retrieval*. 595–604.

[39] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proceedings*

of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 555–562.

[40] Andrew Turpin and Falk Scholer. 2006. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 11–18.

[41] Mengting Wan, Di Wang, Matt Goldman, Matt Taddy, Justin Rao, Jie Liu, Dimitrios Lymberopoulos, and Julian McAuley. 2017. Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs. In *Proceedings of the 26th International Conference on World Wide Web*. 1103–1112.

[42] Xiaojie Wang, Zhicheng Dou, Tetsuya Sakai, and Ji-Rong Wen. 2016. Evaluating search result diversity using intent hierarchies. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 415–424.

[43] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. 2018. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 610–618.

[44] Xiaojie Wang, Jianzhong Qi, Kotagiri Ramamohanarao, Yu Sun, Bo Li, and Rui Zhang. 2018. A joint optimization approach for personalized recommendation diversification. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22*. 597–609.

[45] Xiaojie Wang, Ji-Rong Wen, Zhicheng Dou, Tetsuya Sakai, and Rui Zhang. 2017. Search result diversity evaluation based on intent hierarchies. *IEEE Transactions on Knowledge and Data Engineering* 30, 1 (2017), 156–169.

[46] Edwin B Wilson. 1927. Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.* 22, 158 (1927), 209–212.

[47] Liang Wu, Diane Hu, Liangjie Hong, and Huan Liu. 2018. Turning clicks into purchases: Revenue optimization for product search in e-commerce. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 365–374.

[48] Emine Yilmaz, Javed A Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 587–594.

[49] Fan Zhang, Yiqun Liu, Jiaxin Mao, Min Zhang, and Shaoping Ma. 2020. User behavior modeling for web search evaluation. *AI Open* 1 (2020), 40–56.

[50] Zhen Zuo, Lixi Wang, Michinari Momma, Wenbo Wang, Yikai Ni, Jianfeng Lin, and Yi Sun. 2020. A flexible large-scale similar product identification system in e-commerce. In *KDD Workshop on Industrial Recommendation Systems*.