Real-Time Personalized Ranking in E-commerce Search

Lucia Yu*, Ethan Benjamin*, Congzhe Su†, Yinlin Fu†, Jon Eskreis-Winkler†,
Xiaoting Zhao, Diane Hu
Etsy, Inc
New York, U.S.A
{lyu,ebenjamin,csu,yfu,jeskreiswinkler,xzhao,dhu}@etsy.com

ABSTRACT

Marketplace platforms offer the convenience of browsing through entire catalogs of offerings via a search bar. At Etsy[1], an unconventional inventory of over 100 million unique handmade products presents even greater challenges in product search given that many listings fall outside of standard categories. Personalizing the search experience becomes increasingly important with the overwhelming number of relevant items per query, especially for top percentile of queries by search volume. In accounting for individual user preferences, a personalized model sifts through millions of listings to find a vital few that match their intent and taste.

In this paper, we show how we use a combination of content-based, graph-based and session-based listing representations to construct user and query representations from multiple implicit feedback types aggregated over various time frames to build a personalized learning-to-rank model at Etsy. With rigorous offline evaluations and three online A/B tests conducted across platforms, we show that the proposed personalized ranking variants significantly outperform the existing baselines in ranking metrics with measurably higher degrees of personalization given by Kendall Tau correlation coefficient. Since 2020, the personalized models have been successfully deployed on live traffic at scale across many platforms, with users converting more and faster with higher repeated repurchases. We also provide a deeper analysis of various segments (i.e, query or user bins, platforms) to uncover where personalization shines.

KEYWORDS

Personalization; E-commerce; Search Ranking

1 INTRODUCTION

At Etsy, an unconventional inventory of over 100 million unique handmade products can become overwhelming for buyers to find what they are looking for. Etsy's product search is one common way for buyers to browse through the catalog, and personalizing this search experience thus becomes important in helping the user find items that best fit their preference, as seen in Figure 1.

Many of the most popular queries (also referred to "top" or "head" queries; see Table 2 for details) are broad and short in length. For the query "necklaces", Etsy offers over 300k listings that fall into this product category. Some example items that fall into this group are lockets, chains, and personalized nameplate necklaces. This, and other similar head queries that are vague in nature, are not only the most searched queries, but also the most purchased. In 2020, the top searched query on Etsy was "personalized gifts", which had over 5 million search results [21], as each buyer might look for different products for the gifts.

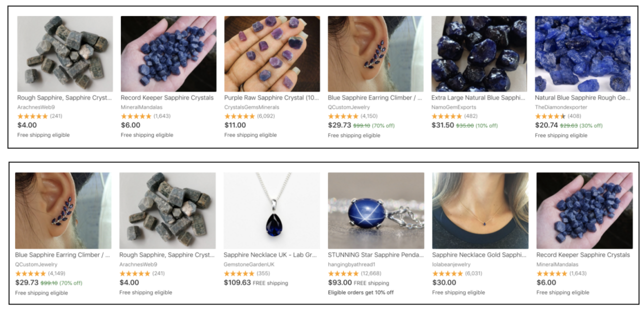

Figure 1: Personalized results for two different users given search query "sapphire". Top user previously purchased "gemstone", "crystal", and "birthstone" items. Bottom user purchased "necklace" and "jewelry" items.

Tail queries such as "woolen upcycle coat" and "early renaissance canvas print" tend to be specific and lengthier than head queries. Although one might argue that personalization is less important for tail queries (since there are potentially fewer listings that fit the query) we show that our personalized model with user profile and query representations improves conversion rates on tail queries also.

Intuitively, the more implicit feedback a user provides in the form of clicks or purchases the better our model can learn their preferences. However we show that our personalized model performs well even on user segments that haven't often interacted with the marketplace. For sellers too, personalization might help the model better recognize the unique qualities of their listings and find the best suited buyers.

Personalization has been shown to improve the user experience and increase the relevancy of returned results [1]. There are many different approaches to personalization but at the center of it lies the user profile, such as those described in [7]. Motivated by these works, our personalized model focuses on behavior modeling using user implicit feedback representations. The user embeddings are aggregated listing representations, of various types. In summary, we discuss the personalized search ranking model and analyze the different features that lead to an improved user experience and increased conversion.Our contributions are as follows:

- We use content-based, graph-based and session-based listing embeddings for personalization to build individual user profiles. We generate these embeddings from four modes of user implicit feedback (clicks, purchases, favorites, cart adds), combined with various time frames of aggregation (i.e, recent vs lifetime), to improve our ranking model relevancy metrics.
- We show that this personalization model deployed on live traffic improves user rates of return and conversion rate. We

[1] E-commerce platform for handmade products at https://www.etsy.com

* Authors have equally contributed to this work.
† Authors have equally contributed to this work.

also share the insights on how they perform on various user and query segments.
- We measure the degree to which our model personalizes search results. Head queries' results as well as traffics on the mobile application platform (where habitual users reside) have lower average Kendall Tau correlation coefficient and thus higher degree of personalization than tail queries.

2 RELATED WORKS

Research on personalization within Search has garnered increased interest in recent years. In particular, usage of implicit feedback logs to generate various embeddings have proven successful in real-world applications [8, 14, 26]. Research on user profiles have explored different ways to create representations from implicit feedback signals. Clicks and purchases are among some of the more common signals, but add-to-collection and favorites signals also reveal user interests and taste preferences. A few approaches to creating such embeddings include the work in [3, 8, 14, 18, 28, 30], as well as topic models to build user profiles for personalization in [6, 12, 20, 25].

With learned embeddings, user representations can be constructed using in-session implicit feedback [3, 8, 30], or a combination of recent and longer term window ranges [2, 8, 15, 24, 28] for downstream tasks in search and recommendations. Authors in [3] show that combining content-based features and content-agnostic based item embeddings on users' recent clicks can improve mean reciprocal rank in an e-commerce setting. [2, 5] constructs user profiles as a combination of previously purchased items for their zero attention model by applying weights to interactions and recency.

In production, search and recommendation systems are typically comprised of two stages: the first pass narrows down the product catalog to a subset of relevant candidates [13, 17, 23, 29, 30], while the second pass performs finer-grained re-ranking of the items to optimize for relevancy and other business metrics [9, 10, 16, 19, 22, 27]. In the past, personalization was applied to the retrieval step like in [3, 29] or both retrieval and ranking passes demonstrated in [26]. In this paper we apply personalization to the re-ranking step.

[11] proposes the degree of personalization for web search results to examine the difference in rankings between search results among users. We analyze the effect of personalized search results across query segments and platforms to examine the degree to which our model generates different rankings among users for the same query.

3 METHODOLOGY

In the following sections we describe the building blocks of our user profile and query embedding features: listing representations. We then explain how we create multiple profiles per user based on various time windows of aggregation and types of interactions. Finally we discuss feature engineering and detail the underlying re-ranking model used by both the baseline and personalized models.

3.1 Listing and Query Representations

Three main listing representations include: content embeddings, item-interaction embeddings, and interaction-based graph embeddings. The first two are listing specific while the third produces both query and listing embeddings per interaction type. Modeling interactions in the latter two approaches are inspired by the hypothesis that the different ways in which a user interacts with an item indicates different kinds of intent, evidenced in [30].

Content Representations, L_C. To capture listing content features, we construct sparse vectors from titles, tags and other seller-contributed texts using Tf-Idf and BM25. We construct up to trigrams over the corpus of all available listings.

Item Interaction Embeddings, $\{L_{S,i}\}_{i\in\mathcal{I}}$. Similar to [30], we learn item interaction embeddings by using one-year session data with interactions (i.e., click, favorite, purchase) to construct dense vectors that represents co-occurrence patterns with respect to user implicit feedback. These embeddings are trained using a skip-gram model, where an instance of data constructed on a given user session, $\mathcal{S} = \{p_1, ..., p_k\}$, consists a sequence of item-interaction tuples, $p_j \in \mathcal{L}\times\mathcal{I}$, with $\mathcal{L}$ and $\mathcal{I}$ denote the set of items and interaction types. As a result, a click-interaction listing embedding will have a different embedding than a purchase-interaction for the same listing.

Interaction-based Graph Embeddings, $(\{L_{G,i}\}_{i\in\mathcal{I}}, \{Q_{G,i}\}_{i\in\mathcal{I}})$. Inspired by the initial work in [14], we extend the vector propagation algorithm to learn representations per *interaction type* (i.e., click, add-to-cart, purchase) for both queries, $\{Q_{G,i}\}_{i\in\mathcal{I}}$, and listings, $\{L_{G,i}\}_{i\in\mathcal{I}}$, in the same semantic space. For each interaction I, we construct a bipartite graph G_I with the node set $V_I = \{P, Q\}$, where an edge $e \in E_I$ connects a listing, $p \in P$, and a query, $q \in Q$, if there is at least one co-interaction between them. These vectors consider content and user interaction data in a shared semantic space of query and listing vocabulary. We propagate from listings to queries, such that listings with commonly associated queries in the local neighborhood of the graph would have similar vectors to reduce semantic gap. The representations are trained over a year's worth of data on user interaction logs (i.e, clicks) to maximize model quality to account for heavy seasonality effect, thus increasing coverage rate for rare tokens. New listings or queries without interaction data outside the graphs leverage the learned token representation in the vocabulary to achieve nearly full coverage rate. One might argue to build one single bi-partite graph with weighted edges among different interactions, but we observe better improvement with interaction-based approach and also it is efficient at scale.

3.2 User Representations

For a given user, our approach to personalization aggregates representations of listings with which the user engaged to create a user profile based on implicit feedback averaging over different time windows: recent (i.e, last 14 days) and lifetime (i.e, all historical purchases or carts from the users). We use all three types of listing representations described in section 3.1, four modes of implicit feedback (click, favorite, add-to-cart, and purchase), and two time windows for user feedback (recent and lifetime).

For a given user u, let $s_t^k \subset S_u$ be the set of listings that the user engaged in the last window t for interaction type equals k. Let S be the set of all possible permutation for s_t^j. For s_t^k and every listing embedding, we can construct multiple user representations as:

$$\{U_{C,s_t^k}, U_{S,s_t^k}, U_{G,s_t^k},\}_{s_t^k\in S},$$

where $U_{C,s_t^k} = Agg(\cup_{\{m:m\in s_t^k\}} L_C^m)$, for example, is a user representation derived based on the listing content representations.

For instance, we take all the recent listings a user has favorited in the last weeks (i.e, 14 days), retrieve the item interaction embeddings for each of these listings and average the vectors to create a user's item interaction embedding for their recently favorited items.

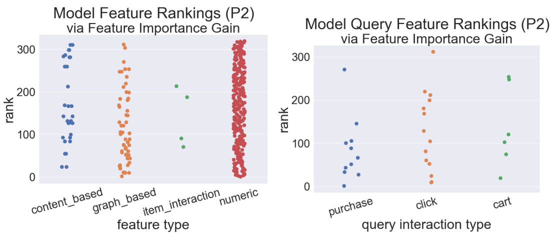

Figure 2: Ranking of features based on feature importance gain for feature types and query interaction types, respectively.

Different time windows for recency have been explored, we finalize to recent clicks and carts over 14 days while considering lifetime favorites and purchases. A more granular approach to weighting user implicit feedback learned through attention mechanisms could be extended, however that is outside the scope of this work.

3.3 Personalized Learning-to-Rank Model

For the baseline and the personalized variants we use an ensemble gradient boosted decision tree with LambdaMART algorithm in the second pass of the information retrieval system [4]. Here is the mathematical formulation. Let q denote the query from a user u and $l_i \in L_{(q,u)}$ denote the i-th listing in the set of listings associated with q and u. For each l_i, let x_i denote the feature vector for this (query, user, listing) tuples, and r_i denote the listing's relevancy (i.e, 1 if the user purchases the listing). Given search data logs $\{q, u, [(x_i, r_i)]_i^{L_{(q,u)}}\}$, we aim to learn a (personalized) ranker, f, that predicts a relevance score on the l given a query q (and user u) through the minimization of the empirical risk function as follows, $\hat{f} = \arg\min_f \sum_{(q,u)} \sum_{l_i \in L_{(q,u)}} Loss\big(f(x_i), r_i\big)$.

For personalized models, we experimented with two variations. The first variation (**P1**) contains user embedding features that the non-personalized model (**B**) doesn't have. Compared to the P1, the second model (**P2**) adds query embedding features learned from the Interaction-based query-listing Bipartite Graph described in Section 3.1. Query features are then engineered to interact with user representations, creating more personalized features for ranking. We experiment with incremental models to show that with each addition of user and query embeddings, we get further model improvements.

3.4 Ranking Features

The baseline model uses both sparse and numeric features that describe listings, shops and query. Some of the raw features include dwell time (i.e, average dwell time per listing), product attributes (i.e, color, material types), taxonomy (i.e, clothing, jewlery), and binned query frequency statistics. We create ratios, normalize and combine composition features from the query to the listings or shops [9, 27]. The personalized models include all the baseline model features plus new features generated with query and user representations detailed in Section 3.1 and 3.2.

The personalized learning-to-rank models receive similarity scores (i.e., cosine, dot product) calculated between user profile or query representations and candidate listing embeddings. For the most part, scores are generated across the same vector types. For example, we generate similarity scores between a user's recent clicks content vector and all candidate listings' content vectors. These inputs are among the few hundred total number of features passed to the decision tree, which include raw features and baseline features.

Models	Purchase NDCG @10		Kendall Tau (avg)	
	Web Traffic	*App Traffic*	*Web Traffic*	*App Traffic*
P1 (user reps)	+3%	+4.8%	0.9073	0.8109
P2 (query + user reps)	+6.9%	+9.17%	0.8527	0.7783

Table 1: Offline evaluations of personalized models P1 and P2 vs Baseline (non personalized) on attributed purchase search requests, measured by % change in NDCG@10 and degree of personalization in Kendall Tau correlation coefficients. A lower Kendall Tau score indicates greater degree of personalization.

4 EXPERIMENTAL RESULTS

In this section, we conduct and discuss the offline and online test results of the base models and personalized variants. Wea also demonstrate that the degree to which results are personalized is measurably different among query segments and across platforms.

4.1 Dataset

We collect training data that consists of purchase search logs from users to the site from over 30 days. That is, when a user confirms a purchase from the site the data contains the query used, purchased item and details about the item such as the tags or taxonomy. The training data also maintains record of the purchased listing in the context of all other listings shown in the results page. In offline experiments cart-add logs were combined with purchase logs to create training data, however we found that training only on purchase logs improved the model performance. Training data with over 200 millions instances was constructed from purchase logs from all platforms on the site, and testing was evaluated on the next day's data and done separately on web and mobile application traffic with three millions testing instances per platform. User traffic on the mobile application falls into the more active category, as they typically tend to be signed-in and have more recurring purchases compared to web traffic users.

4.2 Offline Evaluation and Feature Importance

To evaluate the offline performance of the personalized variants we use purchase NDCG@k, i,e, $k = 10$. In offline experiments shown in Table 1, P1 improved over the baseline model (B) purchase NDCG@10 by 3% and P2 improved over the baseline by 6.9% in purchase NDCG@10 on web traffic. Users that had purchased many items in the past 12 months saw a higher increase in NDCG gain compared to the average user for both P1 and P2 variants in all platforms.

To examine the effect of different vector types, time ranges of user aggregated interaction, and modes of user implicit feedback on purchase NDCG we review the overall rankings of features ordered by the greatest feature importance gains in the tree model. Among embeddings for user profiles, features that included content-based Tf-Idf vectors had higher importance gain, followed by interaction-based graph vectors that learn listing and query neighborhoods defined by co-interactions and graph adjacent neighborhoods. The n-gram token weights of the graph embeddings are chosen by propagating query n-grams and similar listings, whereas Tf-Idf weighs n-gram tokens relative to appearances in the entire listing catalog.

Recent time windows of aggregation for user profiles generally had higher feature importance gain compared to overall time windows. In this case, users' current shopping mission might be more informed by their recent user activity. However the user vectors computed for overall time ranges still remain consequential given their feature gains relative to a randomly generated control feature. This

Query Segments	% Search Traffics	Median Length	Kendall Tau P1 Web	P1 App	P2 Web	P2 App
top 0.01%	>= 99.99%	13	0.873	0.751	0.850	0.719
top 0.1%	>99.90% and <=99.99%	16	0.918	0.859	0.910	0.824
head	>96% and <=99.90%	18	0.970	0.948	0.965	0.909
torso	>70% and <=96%	21	0.995	0.991	0.987	0.946
tail	<=70%	23	0.999	0.993	0.996	0.949

Table 2: The table provides definition of query segments binned by % search volumes, median query length, and degree of personalized results per query segments for P1 and P2 on web and App.

might indicate that lifetime user behavior can still inform general preferences in users' shopping missions.

Across the personalized models, user vectors aggregated across clicks and query vectors aggregated across purchases created features with more feature importance gain compared to cart-adds and favorites. For user profile embeddings, clicks are important signals whereas query embeddings favor purchase signals. Click-based query embeddings have higher coverage in the training data compared to the others, and despite purchases have lower coverage in the training data the signal is stronger than cart-adds.

4.3 Measuring Degree of Personalization

To more deeply understand the effects of personalization, we examine the degree to which the results are personalized for query bins using a method similar to [11]. To measure the degree of personalization for an entire model, we average Kendall Tau coefficients across all queries. We group queries into their bins and average across each bin to obtain the degree of personalization per query segment.

Table 2 and Figure 3 show that the baseline, non-personalized model, generated results for users with the highest degree of similarity of rankings across all queries on a given day from web and app traffic after the training window. Without personalized features, we should expect the results to be the same across users. For the personalized model with user profile embeddings, the average Kendall Tau coefficient decreases and adding query embeddings the coefficient is the lowest. The personalized models are measurably different among different users for each query compared to the baseline. Comparing across platforms, users on our mobile application receive even more personalized results than web traffic. Historically, mobile application users tend to visit and purchase more often than the average web user thus providing the model with more user feedback to generate results with a greater degree of personalization.

The personalized model with user embeddings serves more personalized results for broader, popular queries than for tail queries. We see in Table 2 that the top 0.01% of queries have the lowest Kendall Tau coefficient of all query bins. Tail and torso queries exhibit high Kendall Tau coefficients between users, with their mode around 1.0. Figure 3 plots the Kendall Tau correlation for each model and traffic segment, and shows the modes for top 0.01%, top 0.1% and head queries around less than 0.9 Kendall Tau correlation.

4.4 Online Results

For online A/B testing, we conducted live traffic experiments on all platforms including desktop, mobile web and mobile application over the course of a week, randomly bucketing users into control versus variant with 50/50 split.

In the personalized variants we observe over 3% in increases of purchase NDCG@10, consistent with offline results. The overall

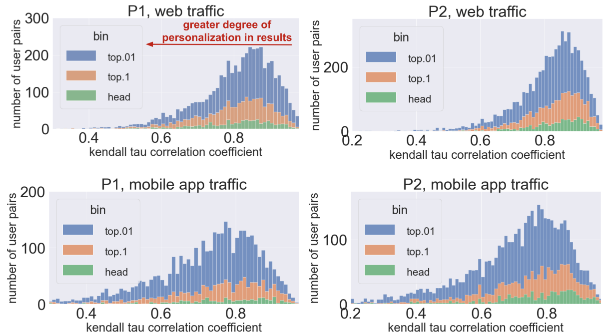

Figure 3: Degree of personalized search results per model for web and mobile application traffic for top queries.

Segments (Metrics in % change)	P1 vs Baseline Web Traffic CVR	CTR	P2 vs P1 Web Traffic CVR	CTR	App Traffic CVR	CTR
Query: top .01%	+0.4%***	+0.81%	+0.23%**	+2.4%*	+0.04%	+11.8%**
Query: top .1%	+.37%**	+1.26%	+0.29%**	+5.6%***	+0.07%	+13.2%*
Query: head	+0.35%***	+1.2%	+0.11%	+4.0%**	+0.22%	+21.0%***
Query: torso	+0.14%	+1.69%	+0.25%	+7.2%**	+0.37%	+27.7%**
Query: tail	+0.13%	−0.32%	+0.71%***	+6.6%*	+1.3%**	+6.4%**
User: habitual	+0.4%*	−1.5%	+0.27%*	+3.3%	+0.2%	+0.26%
User: active	+0.61%**	−2.1%	+0.36%	+3.4%	+0.32%	+11.6%
Overall	+0.65%**	n/a	+0.59%**	n/a	+1.1%**	n/a

Table 3: A/B test results measured by % changes in conversion rates (CVR) and click-through-rate (CTR) for query and user segments: (a) P1 vs baseline (Web), (b) P2 vs P1 (Web), (c) P2 vs P1 (Mobile App). Here, $(*)$, $(**)$, $(***)$ indicate statistical significance at p-value < 0.1, 0.05, 0.01 levels.

user conversion rate increases while the mean search clicks per session decreases in the P1 personalized model compared to baseline. On average, users served this personalized variant purchase more items using fewer number of clicks during the search session. User repurchase rates, or the portion of users who bought a subsequent item within the span of 60 days, also increase.

In P2, query features interact with user profile features to create a contextualized representation of the user's query in addition to the user profile embedding features built on implicit feedback. With these features, online experiments observed further increases in purchase NDCG@10 as well as conversion rate compared to P1.

Users with a purchase within the last 12 months are considered more active users, while all other users are considered less active users. We observe that adding user profile embeddings increases conversion rates for more active users, and adding query embeddings increases conversion rates for less active users. Representing queries via interaction-based graph embeddings helps the model to learn query context, even if the user has a sparse history.

To analyze the conversion rates on different queries, we bin queries into top 0.01%, top 0.1%, head, torso, and tail segments based on search volume over a year, see Table 3.

Adding personalized user profile features in P1 increases conversion rates for the broadest, most popular queries. User profile features also increase the add-to-cart rates for queries in all bins except tail. With the addition of query embeddings in P2, we get a further boost in conversion rates for tail queries too. Contextualized query representations help rarer queries to find suitable listings.

5 CONCLUSION AND FUTURE WORK

In this paper, we discuss how we build personalization in the second pass search ranking via user profile and query representations constructed based on multiple implicit feedback types and various time windows of aggregation. With these features, purchase NDCG@10 and user conversion rates increase overall. Personalization affects users differently, with active users converting at a greater rate due to their richer user history compared to inactive users. The traffic on the mobile application platform generates more personalized results compared to web traffic. We measure the degree to which personalization affects different query segments and found that the top 0.01% of head queries generate the lowest similarity of rankings between users, as measured by the Kendall Tau correlation coefficient.

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2019. Improving Web Search Ranking by Incorporating User Behavior Information. *SIGIR Forum* 52, 2 (Jan. 2019), 11–18. https://doi.org/10.1145/3308774.3308778

[2] Qingyao Ai, Daniel N. Hill, S. V. N. Vishwanathan, and W. Bruce Croft. 2019. A Zero Attention Model for Personalized Product Search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 379–388. https://doi.org/10.1145/3357384.3357980

[3] Grigor Aslanyan, Aritra Mandal, Prathyusha Senthil Kumar, Amit Jaiswal, and Manojkumar Rangasamy Kannadasan. 2020. Personalized Ranking in ECommerce Search. In *Companion Proceedings of the Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 96–97. https://doi.org/10.1145/3366424.3382715

[4] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23-581 (2010), 81.

[5] Mark J. Carman, Fabio Crestani, Morgan Harvey, and Mark Baillie. 2010. Towards Query Log Based Personalization Using Topic Models. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (Toronto, ON, Canada) *(CIKM '10)*. Association for Computing Machinery, New York, NY, USA, 1849–1852. https://doi.org/10.1145/1871437.1871745

[6] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A Large-Scale Evaluation and Analysis of Personalized Search Strategies. In *Proceedings of the 16th International Conference on World Wide Web* (Banff, Alberta, Canada) *(WWW '07)*. Association for Computing Machinery, New York, NY, USA, 581–590. https://doi.org/10.1145/1242572.1242651

[7] Min Gao, Kecheng Liu, and Zhongfu Wu. 2010. *Personalisation in web computing and infromatics: Theories, techniques, applications, and future search*. 607–629.

[8] Mihajlo Grbovic and Haibin Cheng. 2018. Real-Time Personalization Using Embeddings for Search Ranking at Airbnb. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 311–320. https://doi.org/10.1145/3219819.3219885

[9] 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* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 2852–2860. https://doi.org/10.1145/3394486.3403336

[10] Malay Haldar, Prashant Ramanathan, Tyler Sax, Mustafa Abdool, Lanbo Zhang, Aamir Mansawala, Shulin Yang, Bradley Turnbull, and Junshuo Liao. 2020. Improving Deep Learning for Airbnb Search *(KDD '20)*.

[11] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring Personalization of Web Search. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 527–538. https://doi.org/10.1145/2488388.2488435

[12] Morgan Harvey, Fabio Crestani, and Mark J. Carman. 2013. Building User Profiles from Topic Models for Personalised Search. In *Proceedings of the 22nd ACM International Conference on Information Knowledge Management* (San Francisco, California, USA) *(CIKM '13)*. Association for Computing Machinery, New York, NY, USA, 2309–2314. https://doi.org/10.1145/2505515.2505642

[13] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-Based Retrieval in Facebook Search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*.

[14] Shan Jiang, Yuening Hu, Changsung Kang, Tim Daly, Dawei Yin, Yi Chang, and Chengxiang Zhai. 2016. Learning Query and Document Relevance from a Web-Scale Click Graph. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 185–194. https://doi.org/10.1145/2911451.2911531

[15] Henry Lieberman. 1995. Letizia: An Agent That Assists Web Browsing. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1* (Montreal, Quebec, Canada) *(IJCAI'95)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 924–929.

[16] Rishabh Mehrotra, Niannan Xue, and Mounia Lalmas. 2020. Bandit Based Optimization of Multiple Objectives on a Music Streaming Platform. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3224–3233. https://doi.org/10.1145/3394486.3403374

[17] Priyanka Nigam, Yiwei Song, Vijai Mohan, Vihan Lakshman, Weitian (Allen) Ding, Ankit Shingavi, Choon Hui Teo, Hao Gu, and Bing Yin. 2019. Semantic Product Search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*.

[18] Aditya Pal, Chantat Eksombatchai, Yitong Zhou, Bo Zhao, Charles Rosenberg, and Jure Leskovec. 2020. PinnerSage: Multi-Modal User Embedding Framework for Recommendations at Pinterest. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 2311–2320. https://doi.org/10.1145/3394486.3403280

[19] Przemysław Pobrotyn, Tomasz Bartczak, Mikołaj Synowiec, Radosław Białobrzeski, and Jarosław Bojar. 2020. Context-Aware Learning to Rank with Self-Attention. arXiv:2005.10084

[20] Micro Speretta and Susan Gauch. 2005. Personalized Search Based on User Search Histories. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI '05)*. IEEE Computer Society, USA, 622–628. https://doi.org/10.1109/WI.2005.114

[21] Kevin Stankiewicz. 2020. *'Personalized gifts' is the No. 1 search term on Etsy this year, CEO says*. https://www.cnbc.com/2020/11/30/holiday-shopping-personalized-gifts-no-1-search-term-on-etsy-in-2020.html

[22] Andrew Stanton, Akhila Ananthram, and Pablo Crespo. 2020. Evolving Multi-Objective Ranking Models for GMV Optimization in E-Commerce. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion* (Cancún, Mexico) *(GECCO '20)*. Association for Computing Machinery, New York, NY, USA, 289–290. https://doi.org/10.1145/3377929.3389907

[23] Shulong Tan, Zhixin Zhou, Zhaozhuo Xu, and Ping Li. 2020. Fast Item Ranking under Neural Network Based Measures. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20)*.

[24] Yury Ustinovskiy, Gleb Gusev, and Pavel Serdyukov. 2015. An Optimization Framework for Weighting Implicit Relevance Labels for Personalized Web Search. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) *(WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1144–1154. https://doi.org/10.1145/2736277.2741105

[25] Thanh Vu, Dat Quoc Nguyen, Mark Johnson, Dawei Song, and Alistair Willis. 2017. Search Personalization with Embeddings. *Advances in Information Retrieval* (2017), 598–604. https://doi.org/10.1007/978-3-319-56608-5_54

[26] Jizhe Wang, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee. 2018. Billion-Scale Commodity Embedding for E-Commerce Recommendation in Alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 839–848. https://doi.org/10.1145/3219819.3219869

[27] 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* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 365–374. https://doi.org/10.1145/3209978.3209993

[28] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2020. Product Knowledge Graph Embedding for E-Commerce. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 672–680. https://doi.org/10.1145/3336191.3371778

[29] Han Zhang, Songlin Wang, Kang Zhang, Zhiling Tang, Yunjiang Jiang, Yun Xiao, Weipeng Yan, and Wen-Yun Yang. 2020. Towards Personalized and Semantic Retrieval: An End-to-End Solution for E-Commerce Search via Embedding Learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2407–2416. https://doi.org/10.1145/3397271.3401446

[30] Xiaoting Zhao, Raphael Louca, Diane Hu, and Liangjie Hong. 2020. The Difference Between a Click and a Cart-Add: Learning Interaction-Specific Embeddings. In *Companion Proceedings of the Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 454–460. https://doi.org/10.1145/3366424.3386197