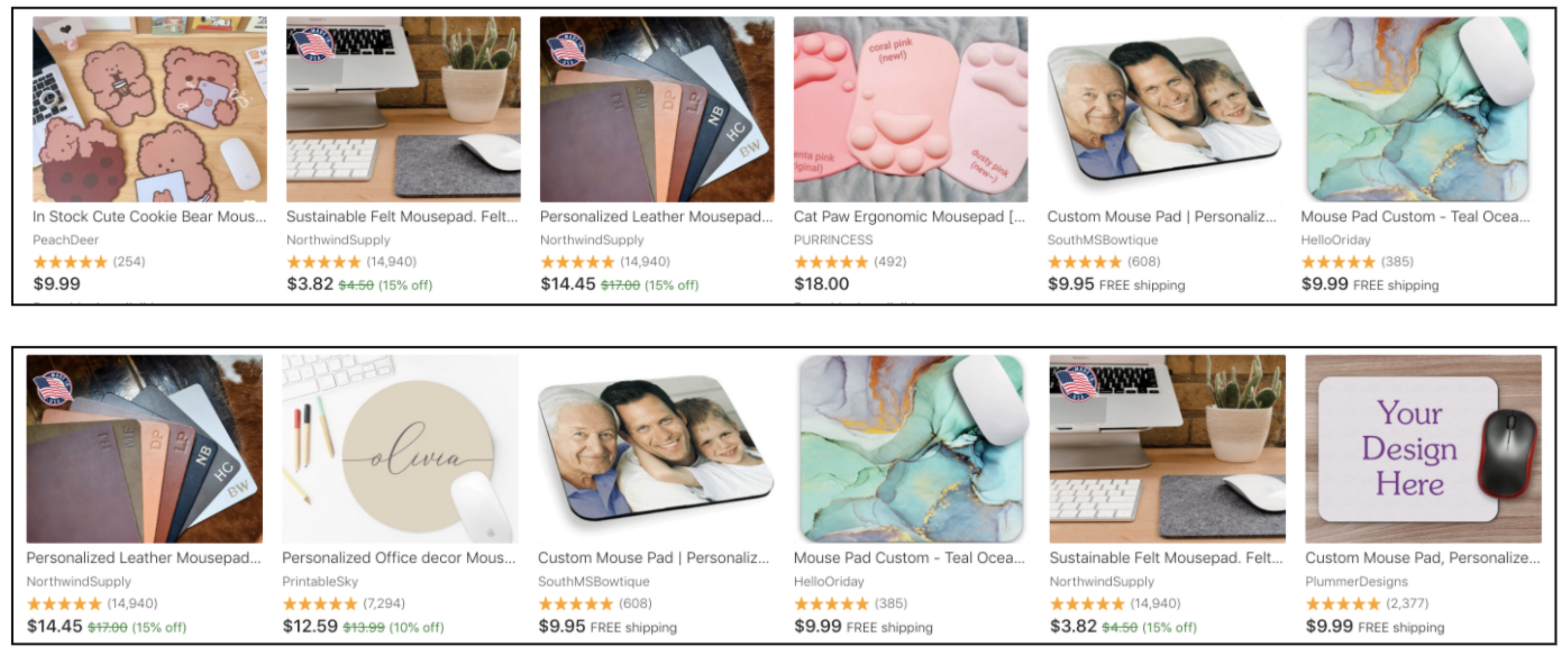
Personalization in E-commerce Product Search by User-Centric Ranking

Lucia Yu   Ethan Benjamin   Congzhe Su   Yinlin Fu   Jon Eskreis-Winkler   Xiaoting Zhao   Diane Hu

{ lyu, ebenjamin, csu, yfu, jeskreiswinkler, xzhao, dhu } @etsy.com

Motivation

With a large pool of candidate listings, personalized search is an important tool to **help users find items that best fit their preferences**. Head queries account for the lion's share of purchases from the site. Many of these queries are vague, short in length, and have a myriad of search results. As an example, in 2020 the top searched query on Etsy was "personalized gifts", which had over 5 million search results.



In Stock Cute Cookie Bear Mous... PeachDeer ★★★★★ (254) $9.99
Sustainable Felt Mousepad. Felt... NorthwindSupply ★★★★★ (14,940) $3.82 ~~$4.50~~ (15% off)
Personalized Leather Mousepad... NorthwindSupply ★★★★★ (14,940) $14.45 ~~$17.00~~ (15% off)
Cat Paw Ergonomic Mousepad [... PURRINCESS ★★★★★ (492) $18.00
Custom Mouse Pad | Personaliz... SouthMSBowtique ★★★★★ (608) $9.95 FREE shipping
Mouse Pad Custom - Teal Ocea... HelloOriday ★★★★★ (385) $9.99 FREE shipping

Personalized Leather Mousepad... NorthwindSupply ★★★★★ (14,940) $14.45 ~~$17.00~~ (15% off)
Personalized Office decor Mous... PrintableSky ★★★★★ (7,294) $12.59 ~~$13.99~~ (10% off)
Custom Mouse Pad | Personaliz... SouthMSBowtique ★★★★★ (608) $9.95 FREE shipping
Mouse Pad Custom - Teal Ocea... HelloOriday ★★★★★ (385) $9.99 FREE shipping
Sustainable Felt Mousepad. Felt... NorthwindSupply ★★★★★ (14,940) $3.82 ~~$4.50~~ (15% off)
Custom Mouse Pad, Personalize... PlummerDesigns ★★★★★ (2,377) $9.99 FREE shipping

Figure 1. Results for query "mousepad". Top user previously purchased "kawaii" and "cute" items; bottom user purchased "monogram" and "family photo" items.

Contributions

1. **User-Centric Ranking**. 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.
2. **Measurably Personalized Results**. The personalized ranking variants have measurably higher degrees of personalization, given by Kendall Tau correlation coefficients. The mobile application platform, characterized by a higher number of active users, demonstrates higher levels of personalization compared to web traffic. The top 0.01% of queries exhibit the highest degree of personalization compared to torso and tail bins.

Models

We use an ensemble **gradient boosted decision tree with LambdaMART algorithm** in the second pass of the information retrieval system. We experiment with **two personalized variants** in addition to the baseline to track the incremental changes with the addition of new feature groups.

- **Baseline Model** contains non-personalized features such as listing dwell time, listing price and number of shop views collected from purchase logs.
- **P1 Model** includes all the baseline model features, plus user profile representations generated from aggregating three types of listing representations: Tf-Idf, item interaction embeddings, and interaction-based graph embeddings.
- **P2 Model** has all the features in the baseline and P1 model, plus query representations engineered to interact with user representations for an even greater number of personalization features.

Dataset

We re-train our model daily on over 30 days worth of **user purchase logs** collected from our **web, mobile web and mobile app platforms**. A purchase log contains information about the user, the purchased listing, and features related to the user-entered query. Each time we train the model, it sees over **200 million** user purchase logs. A day's purchase log can contain millions of unique users, listings and queries. For testing, we look at data with respect to each platform to measure their performance with the variants.

Evaluation Metrics

For offline evaluation, we look at **purchase NDCG@10**. For online evaluation, we look at purchase NDCG as well as **user conversion rates**, **user clicks per session** and **user rates of return**. To measure the degree of personalization, we look at **Kendall Tau correlation** for query bins, platform traffic, and model variants.

Models	NDCG $@10$		Kendall Tau	
	Web	*App*	*Web*	*App*
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), measured by % change in NDCG@10 and degree of personalization in Kendall Tau coefficients. A lower Kendall Tau score means greater degree of personalization.

Ranking Features

In total, our decision tree used **hundreds of features** to learn to rank. For a given user u, let $s_t^k \subset S_u$ be the set of listings that 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:

$$\{U_{CEmb,s_t^k}, U_{SEmb,is_t^k}, U_{GEmb,s_t^k}, \}_{s_t^k \in S},$$

where $U_{CEmb,s_t^k} = Agg(\cup_{\{m:m\in s_t^k\}} L^m_{CEmb})$, for example, is derived from listing content embeddings.

- **Baseline Features** We create ratios, normalize and combine composition features from query to the listings or shops. Some of these features include historical number of shop purchases and listing dwell time.
- **User Profile Features** We aggregate listing embeddings to create user profile features. We aggregate over user implicit feedback type and recent or lifetime time frames.
- **Query Features** We engineered query features together with user features to create even more personalized features. For example, we get similarity scores of the query interaction-based graph representations with user graph representation and give it as a model input.

Type of listing reps	Type of implicit feedback	window
Tf-Idf	click*	recent
Interaction-based graph vector	cart-add*	lifetime
Item-interaction embedding	favorite	
	purchase	

Table 2. Table of possible user profile feature compositions. * indicates only "recent" time frame for these features

Online Experiment Results

- **Significant CVR on head queries with user representation.** When comparing P1 to baseline, we see significantly higher CVR on web traffic for all three of the top head query bins. Although head query lengths tend to be shorter (see Table 4), this potentially creates opportunity for personalization to really shine.
- **Significant CVR on tail queries with query representation.** With the addition of query representations, we observe significant CVR on tail queries for web traffic on P2 compared to P1. For less-searched queries, interaction-based graph embeddings provide more context and increase the quality of search results.
- **Active users benefit more from personalization.** Mobile app traffic typically has more active users than web and mobile web users. With more more implicit feedback, the CVR for mobile app users was significantly higher overall in P2 compared to P1.
- **Less searching, more buying.** Overall user conversion rate increases while the mean search clicks per session decreases, indicating that users are finding what they want faster. User repurchase rates, or the portion of users who bought a subsequent item within the span of 60 days, increase with personalization.

Segments	P1 vs Baseline		P2 vs P1			
	Web Traffic		Web Traffic		App Traffic	
(Metrics in % change)	*CVR*	*CTR*	*CVR*	*CTR*	*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.

Degree of Personalization

- **Greater personalization for active users.** The more activity a user has, the more personalized their search experience is. Looking at different platforms (web vs. mobile application), more personalized search results are served to mobile application users. Mobile app users tend to be more active users, with more implicit feedback and higher purchase rates.
- **Greater personalization for head queries.** We also see that head queries had more personalized search results than tail queries, with lower Kendall Tau correlation scores.

Query Segments	% Search Traffics	Median Length	Kendall Tau			
			P1		P2	
			Web	App	Web	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 4. 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.

The Web Conference 2021, Knowledge Management in e-Commerce Workshop