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
Etsy, Inc
New York, U.S.A
{lyu,ebenjamin,csu,yfu,jeskreiswinkler,xzhao,dhu}@etsy.com

ABSTRACT

E-commerce platforms offer the convenience of browsing through an entire catalog of inventory via a search bar. An unconventional inventory of unique products presents even greater challenges for product search, given that many of listings fall outside of standard e-commerce categories. With the potentially overwhelming number of relevant items per query, it becomes increasingly important for marketplaces and platforms to help the user find items that best fit their preference and interest via a user-centric ranking model that generates personalized search results. In this paper, we demonstrate how we use a combination of learned content-based and session-based listing representations to build user profiles from multiple implicit feedback types aggregated over various time frames in order to create a personalized tree-based ranking model at Etsy[1]. Etsy is one of largest e-commerce marketplaces with millions of unique, handcrafted items being sold to shoppers around the world. In the proposed personalized model, we observe offline improvements in ranking metrics (i.e, purchase NDCG@10) and higher degrees of personalization measured by Kendall Tau coefficients, when compared to a non-personalized ranking model. We successfully deploy the user-centric ranking model across multiple platforms on live traffic to hundreds of millions of users, with thousands of search requests per second. With the results from three different online A/B experiments, we show that users spend less time searching and buy more items in the personalized variants compared to the baseline.

KEYWORDS

Personalization; E-commerce; Search Ranking

1 INTRODUCTION

E-commerce marketplaces match users with items from sellers that are most relevant to their search query. To give buyers a better search experience, we sift through millions of listings on their behalf to narrow down their search for the perfect item. With a large pool of candidate listings, personalized search results are an important tool to help users find items that best fit their preference, as demonstrated in the search results shown in Figure 1. Many of the most-searched queries during user sessions are broad and short in length. For example, Etsy offers over 300k listings that fall into the "necklaces" category. In 2020 the top searched query on Etsy was "personalized gifts", which had over 5 million search results [17]. Head queries are popular among users, and they account for many of the purchases from the site.

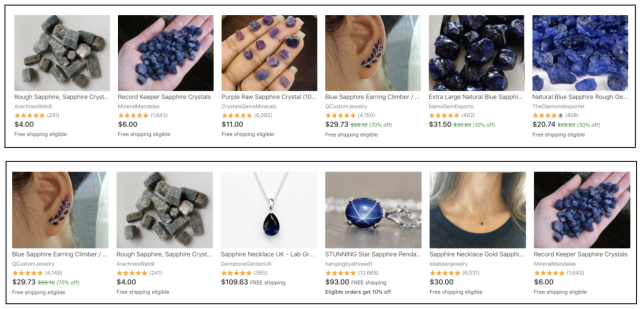

Figure 1: Personalized search results for query "sapphire". User purchased "gemstone", "crystal", and "birthstone" items (top). User purchased "necklace" and "jewelry" items (bottom).

At Etsy, users also search for creative tail queries such as "woolen upcycle coat" and "early renaissance canvas print". These tend to be specific and lengthier than head queries. Although one might argue that personalization is less effective for tail queries (since there might be fewer listings that fit the query), we show that our personalized model with user profile and query representations improves conversion rates on this query bin too.

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.

Personalization has been shown to improve user experience and increase the relevancy of returned results [1]. There exist many approaches to personalization, but at the center of it lies the user profile. [7] define a few types of user profile modeling. Motivated by these works, our personalized model focuses on behavior modeling using user implicit feedback representations. In summary, we discuss the personalized search ranking model and analyze the different features that lead to an improved user experience and overall conversion rate for purchases. Compared to other works on personalization in e-commerce, our contributions are as follows:

- We use both content-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.

[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.

- We also show that the personalization model benefits users with varying levels of historical activity and implicit feedback. We demonstrate that for most queries the personalized model with user and query embeddings improves conversion rates.
- We also measure the degree to which our model personalizes search results. Head queries' results have lower average Kendall Tau correlation coefficient and thus higher degree of personalization than tail queries.

2 RELATED WORKS

There has been a growth of literature that studies different types of user interactions across various marketplaces and platforms to create multiple embeddings, since users often have various ways of providing implicit feedback. Clicks and purchases are common ones, while add-to-collection and favorites can often reveal user interests and taste preferences even if they don't necessarily purchase the item. Approaches to creating such embeddings include the works in [3, 8, 13, 15, 23], as well as topic models to build user profiles for personalization in [5, 12, 16, 19].

With learned embeddings, user representations can be constructed using in-session implicit feedback [3], or a combination of recent and longer term window ranges [2, 9, 14, 18] for downstream tasks in ranking and personalization. Typically, ranking systems are comprised of two stages: the first pass narrows down the product catalog to a subset of relevant candidates, while the second pass performs finer-grained re-ranking of items to optimize for relevancy and other business metrics. Other personalization work has applied personalization to the retrieval step only like in [3, 22], or both retrieval and ranking demonstrated in [20]. In this paper we apply personalization to the re-ranking step. 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, 4] constructs user profiles as a combination of previously purchased items for their zero attention model by applying weights to interactions and recency.

The work in [11] describes a method of measuring the degree of personalization for web search page 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 for different users.

3 METHODOLOGY

In this section, 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 explain our feature engineering and detail the underlying re-ranking model used by both the baseline and personalized models in our experiments.

3.1 Listing Representations

Three main listing representations include: term frequency–inverse document frequency (Tf-Idf), interaction-based graph vectors [13], and item-interaction embeddings [23].

Tf-Idf, extended to BM25, is a content-based sparse vector representation that uses titles, tags and other seller-contributed textual content for the listing with a nearly full coverage rate. We construct up to trigrams over the corpus of all available listings.

Inspired by [13], the interaction-based graph representations propagate listing and query content in a bipartite graph across different session. Vector representations that consider content and session information in a shared semantic space of query and listing vocabulary are learned for listings. We propagate from listings to queries via each *interaction type* (i.e., clicks, add-to-carts, or purchases), such that listings with commonly associated queries in the local neighborhood of the graph would have similar vectors. The representations are trained over a year's worth of data on user interaction logs (i.e, clicks) to increase model quality and coverage rate for rare tokens. New listings not included the graphs can leverage the learned token representation in this large vocabulary.

Item interaction embeddings use interaction types (i.e., click, favorite, add-to-cart, purchase) and session data to construct dense vectors that represent co-occurrence patterns with respect to user implicit feedback. These embeddings are trained using a skip-gram model, where an instance of data is a sequence of (listing ID, interaction type) tuples. As a result a click-interaction listing embedding will be different from a purchase-interaction embedding for the same listing. [23] shows that including item interaction embeddings for a listing can accurately model user shopping behaviors. We train item interaction embeddings over a year's worth of user logs.

3.2 User and Query Representations

For a given user, our approach to personalization aggregates representations of listings on which the user has engaged in order create a user profile based on implicit feedback over different time windows. We use all three types of listing representations and four modes of implicit feedback. User feedback is aggregated over two time windows, recent and lifetime. For example, we take all the recent listings a user has favorited, retrieve the item interaction embeddings for each of these listings and average the dense vectors to create a user's item interaction embedding for their recently favorited items. 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.

In a similar fashion to user representations, we create query representations. Using the interaction-based graph via clicks, purchases and cart adds we average the dense vectors over recent and overall time ranges to represent queries. For example, to generate click-based graph embeddings for a query, we aggregate the embeddings of commonly clicked listings of the query over some time period.

3.3 Learning-to-Rank Model

For the non-personalized baseline and the personalized models we use an ensemble gradient boosted decision tree with LambdaMART algorithm as the second-pass reranking in our product search. The models are trained over a month of implicit purchase logs collected from users to the site. That is, when a user confirms a purchase from the site the data contains the query used, the purchased item and details about the item such as the tags or taxonomy, more in Section 3.4. The training data maintains record of the purchased listing in the context of all other listings shown in the results page. A single day of purchase logs includes millions of requests from millions of unique users, consisting of millions of unique listings.

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

The training duration lasts over 7 hours on 96 vCPUs, and is retrained daily.

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 first personalized model, the second model (**P2**) adds query embeddings features that are feature engineered to interact with user representations to create further personalized features for ranking in P2. We experiment with incremental models to show that a personalized model that can generalize a user's query via query embedding information improves upon a model without these features.

We conduct hyperparameter tuning to find the optimal learning rate, number of trees, data sub-sampling fraction and other model parameters. The final hyperparameter options for P2 used a learning rate of 0.25, 1,500 trees, 0.8 data sub-sampling fraction and a minimum of 100 data samples per leaf.

3.4 Ranking Features

The non-personalized baseline model uses both sparse and numeric features that describe listings, shops and queries. Some of the raw features include dwell time, product attributes, taxonomy information, and binned query frequency statistics. We create ratios, normalize and combine composition features from query to the listings or shops [10, 21]. The personalized models uses all the baseline model features, plus features generated with query and user representations illustrated in Table 1.

For the personalized models, we input to the tree similarity scores between user profile or query representations and candidate listing embeddings, generated across same vector types plus additional compositions. For example, we generate similarity scores between a user's recent clicks Tf-Idf vector and all candidate listings' Tf-Idf vectors. We then pass this score as one input of hundreds to the decision tree, in addition to the baseline features. In offline experiments we compute similarity scores that connect user behavior or query to the candidate listings.

3.5 Production Deployment

In the baseline, query and listing results are cached for reuse between user sessions with an appropriate time-to-live to balance model freshness and latency. This saves costs associated with repeated computations when serving results for non-personalized models. But given that listing results are user specific, a cache can still be implemented however its usage will see a steep decline. To successfully deploy the personalized model into the production, our infrastructure scales up to accommodate for a nearly 20% increase in capacity load.

4 EXPERIMENTAL RESULTS

To evaluate the offline performance of the personalized variants we use purchase NDCG@k, i.e. $k = 10$. In offline experiments cart-add logs were combined with purchase logs to create training data, however we found that training on purchase logs only improved the model performance. Training data used logs from all platforms on the site, and testing was done separately on web and mobile application traffic.

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 2: 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, greater degree of personalization.

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. We bucket a 50/50 split of control versus variant randomly. We examine purchase NDCG@10, conversion rates, user rates of return, and the amount of time users spend searching during their sessions to evaluate the performance of the personalized models.

4.1 Offline Evaluation

In offline experiments shown in Table 2, P1 improved over the baseline model purchase NDCG by 3% and P2 improved over P1 by 3.8% (or P2 improved over Baseline by 6.9%) purchase NDCG. Users that had purchased many items in the past 12 months saw a higher increased 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 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 having the lowest coverage in the training data, purchase signal is stronger than cart adds.

4.2 Measuring Degree of Personalization

To more deeply understand the specific effects of personalization, we examine the degree to which the results are personalized for query bins using a method similar to [11]. The degree of personalization is measured by averaging the Kendall Tau correlation coefficient between all possible combinations of user pairs' search results per

Why do we still measure only NE and AUC for our ranking models?
Is it because our training logs are heavely subsampled?

Query Bins (Segments)	% Traffics (Query Volumes)	Length (Median)	Kendall Tau (Average)
top 0.01%	>= 99.99%	13	0.873
top 0.1%	>99.90% and <=99.99%	16	0.918
head	>96% and <=99.90%	18	0.970
torso	>70% and <=96%	21	0.995
tail	<=70%	23	0.999

Table 3: Degree of personalized results per query bin for P1 model.

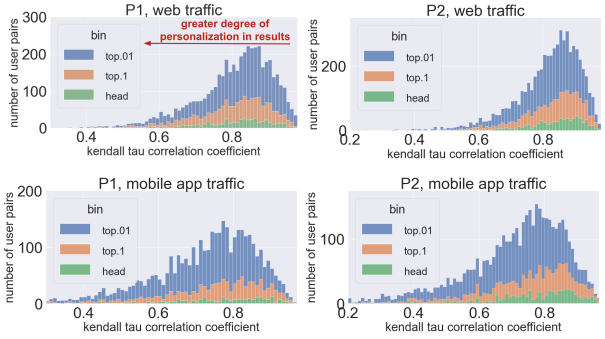

Figure 2: Degree of personalized results per model for web and mobile application traffic for top queries.

given query. To measure the degree of personalization for an entire model, we average Kendall Tau coefficients across all queries. We group queries into their respective bins and average across each bin to obtain the degree of personalization per query segment.

Table 2 and Figure 2 show that the non-personalized, baseline model generated results for users with the highest degree of similarity of rankings across all queries on a given day from web traffic. 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. By adding query embeddings, the coefficient is the lowest compared to other models. 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 with a lower Kendall Tau correlation. 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 3 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 2 plots the Kendall Tau correlation for each model and traffic segment, showing the modes for top 0.01%, top 0.1% and head queries to be less than 0.9 Kendall Tau correlation.

4.3 Online Results

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

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%	.4%***	.81%	.23%**	2.4%*	.04%	11.8%**
Query: top .1%	.37%**	1.26%	.29%**	5.6%***	.07%	13.2%*
Query: head	.35%***	1.2%	.11%	4.0%**	.22%	21.0%***
Query: torso	.14%	1.69%	.25%	7.2%**	.37%	27.7%**
Query: tail	.13%	-.32%	.71%***	6.6%*	1.3%**	6.4%**
User: habitual	.4%*	-1.5%	.27%*	3.3%	.2%	.26%
User: active	.61%**	-2.1%	.36%	3.4%	.32%	11.6%
Overall	.65%**	n/a	.59%**	n/a	1.1%**	n/a

Table 4: 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.

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 4.

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 find suitable listings.

5 CONCLUSION AND FUTURE WORK

In this paper, we discuss how we build personalization in the second pass of product search 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.

For search results pages, there can be a positional biases of the listings being showed in grid layout of results [10]. Future work should work to account for these positional biases during the re-ranking step. Fairness and inclusivity is an important component of search ranking and recommendations. [6] develops granular metrics to accurately assess model biases and corrects them with machine learning techniques. Listings should be ranked such that relevant listings are properly represented while mitigating forms of visual bias or otherwise.

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] 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

[5] 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

[6] Nadia Fawaz. 2020. Powering inclusive search recommendations with our new visual skin tone model. https://medium.com/pinterest-engineering/powering-inclusive-search-recommendations-with-our-new-visual-skin-tone-model-1d3ba6eeffc7

[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] Mihajlo Grbovic and Haibin Cheng. 2018. *Real-Time Personalization Using Embeddings for Search Ranking at Airbnb*. Association for Computing Machinery, New York, NY, USA, 311–320. https://doi.org/10.1145/3219819.3219885

[10] 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

[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] 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

[14] 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.

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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

Main ideas:

* User-emb – average of items emb. (over different time windows and interaction types)
* Item emb – TF-IDF, item-interaction emb.
* Query emb – avg. of items emb.
* Add emb. and sim scores as features to ranking models
* Use kendall-tau to measure degree of personalization
  Measure for different buckets of queries ⇒ more personalization even for tail queries. High degree of pers. for most popular queries.
* Most active users benefits the most!