Long or Short or Both? An Exploration on Lookback Time Windows of Behavioral Features in Product Search Ranking

Qi Liu, Atul Singh, Jingbo Liu, Cun Mu, Zheng Yan and Jan Pedersen

Walmart Global Tech, Hoboken, NJ, U.S.A.

Abstract

Customer shopping behavioral features are core to product search ranking models in eCommerce. In this paper, we investigate the effect of lookback time windows when aggregating these features at the (query, product) level over history. By studying the pros and cons of using long and short time windows, we propose a novel approach to integrating these historical behavioral features of different time windows. In particular, we address the criticality of using query-level vertical signals in ranking models to effectively aggregate all information from different behavioral features. Anecdotal evidence for the proposed approach is also provided using live product search traffic on *Walmart.com*.

Keywords

Online shopping, product search ranking, learning to rank, feature engineering, behavioral features

1. Introduction

Online shopping has become an indispensable part of people's daily lives due to its convenience, wide selection, cost-effectiveness, and mobile accessibility. With an ever increasing catalog size, product search ranking system [1, 2, 3, 4, 5, 6] has been playing a pivotal role in serving customers by ranking relevant products at the top of their search results.

At the heart of every modern eCommerce product search ranking system lies a machine-learned ranking model. For example, LambdaRank/MART [7, 8] leverages gradient boosting machines [9], and neural ranker [10] employs deep learning techniques. These models evaluate and assign scores to each product based on a wide range of input signals derived from diverse sources, including user behaviors, query intents, product attributes, seller reputations, and sophisticated interactions among them.

Out of these many hundreds or even thousands of signals, behavioral features hold significant importance as they are generated through direct interactions between customers and products, encompassing actions like impressions, clicks, add-to-carts (ATCs), purchases, and others. Several studies [1, 11, 12, 13, 14] have emphasized the pivotal role of such implicit relevance feedback [15] in product ranking. In the eCommerce context, customers are the ultimate authorities in determining the relevance of products for a given query, particularly when their judgment is backed by their purchasing decisions. Moreover, such logged customer feedback is

abundant and cheap to obtain in nowadays operational systems. Hence, it is very natural that (query, product)-level behavioral features are the ones that ranking models rely on the most when ranking products.

In spite of the rich and growing literature on leveraging (query, product)-level behavioral features and their variants in product search ranking, one much unaddressed problem is *what the lookback time window should be used to aggregate the customer engagement at the (query, product) level*. This is a very critical design question for all practitioners when applying these essential behavioral features in their ranking systems. In this paper, we will share our empirical insights from our first-hand industrial experience. In particular, we explore behavioral features aggregated over different lookback time windows and study their respective effects on product search ranking. Based upon the pros and cons of using long and short time windows, we propose a principled approach to integrating both sets of behavioral features into the model. The effectiveness of this hybrid model is justified on real product search traffic at *Walmart.com* through online A/B tests.

The remainder of the paper is organized as follows. Section 2 discusses the pros and cons of using behavioral features with long and/or short windows. Section 3 details the proposed enhancement to achieve the best integration of both long and short windows. Section 4 describes the comprehensive online A/B experiment conducted to evaluate the proposed ranking model. Finally, Section 5 summarizes our findings and draws conclusions.

2. Long or Short or Both?

In this section, we will explore the effect of different lookback time window lengths when leveraging behavioral features in product search ranking models. Three types of (query, product)-level user engagement are considered: click rate, add-to-cart (ATC) rate, and order rate. To compute these rates, for a given query (q) - product (p) pair (q, p), we employ the Beta-Binomial Bayesian model and derive behavioral feature values $br_{q,p}$ as the posterior mean of the following Beta distribution,

$$\mathbf{Beta}\left(\sum_{t \in T} b_{q,p}^{(t)} + \alpha, \sum_{t \in T} e_{q,p}^{(t)} - \sum_{t \in T} b_{q,p}^{(t)} + \beta \right), \tag{1}$$

where α and β specify the prior Beta distribution, $b_{q,p}^{(t)}$ is the raw count of the behavior (clicks, ATCs, or orders) frequency for (q, p) on day t, $e_{q,p}^{(t)}$ is the raw count of customer examines for (q, p) on day t, and T is the collection of lookback dates we use to aggregate the engagement data. In particular, the following behavioral features are output to our ranking model,

$$br_{q,p} = \frac{\sum_{t \in T} b_{q,p}^{(t)} + \alpha}{\sum_{t \in T} e_{q,p}^{(t)} + \alpha + \beta}, \tag{2}$$

which is quite similar to the behavioral features defined in [16] but smoothed with prior in order to better address the cold start problem [13].

As shown in equation 2, one critical factor influencing the values and interpretation of behavioral features is the lookback time window length $|T|$ used to aggregate engagements. Utilizing a longer time window captures long-term customer engagement patterns but may

Window Lengths	Long	Short
Pros	• rich in historical engagement data • robust to noise	• good at capturing recent behavioral changes from customers • friendly to new products
Cons	• insensitive to recent behavioral changes from customers • frictional to new products	• sparse in historical engagement data • prone to noise

Table 1
Pros and cons of long and short time windows for behavioral features. The longer the time window, the more historical engagement observations the features capture, which leads to less sparsity and higher coverage in model training and inference. The shorter the time window, the more it captures recent online shopping trends due to customer behavioral changes, new product launches, etc.

overlook recent trends. Conversely, a short time window highlights more short-term behaviors but may not accurately reflect enduring customer interests. Both long and short time windows for aggregating behavioral features present distinct advantages and disadvantages outlined in Table 1.

To investigate the impacts of long- and short-term behavioral features, we define 2 years ($|T| = 730$) as the long lookback time window and 1 month ($|T| = 30$) as the short one, and we specify the ranking model with only 2-year behavioral features as the baseline. Three distinct ranking models with different designs on the window lengths are proposed below.

- **Baseline Model:** model with only 2-year behavioral features.
- **Model A:** model with only 1-month behavioral features.
- **Model B:** model with both 2-year and 1-month behavioral features.

Our search ranking models are trained using XGBoost [17] with the Learning-to-Rank (LTR) framework [18] very similar to [16] by utilizing data from a truncated historical period of online customer search traffic on *Walmart.com* for model training.

To explore the best usage of lookback time windows, we conducted multiple interleaving tests [19], each comparing one proposed model against the baseline model. These tests were performed on a substantial volume of online customer traffic to compare their reactions to different ranking models. Specifically, for each test, we compare the percentage of searches that result in customer engagements between the Control and Variant groups using their respective ranking models. The results are further segmented by different verticals—specific business niches tailored to particular shopping needs. We currently categorize our search queries into six verticals: Food, Consumables, Home, Hardlines, Fashion, and ETS (Electronics, Toys, and Seasonal), with the latter four collectively categorized as General Merchandise (GM).

2.1. Only Long / Short Window

The first interleaving test is configured as follows:

- **Control:** Baseline Model (2-year behavioral features only),
- **Variant:** Model A (1-month behavioral features only),

Vertical	Food	Consumables	Home	ETS	Hardlines	Fashion	Overall
Change	−0.63%*	−0.67%	−0.34%	+3.79%*	+1.51%	+1.06%	−0.28%

Table 2
Result of % searches with engagement for Test 1. The control is the baseline model (only using 2-year behavioral features), and the variant is Model A (only using 1-month behavioral features). The significance level of all tests throughout this paper is set to be 0.1, and the statistically significant results are starred in the tables.

with the purpose to separately examine and compare the individual impact of 2-year and 1-month behavioral features on search ranking models.

The test result is presented in Table 2. Although Model A demonstrates an overall insignificant decline compared to the baseline, when zooming into each business vertical, we find very interesting stories. Model A exhibits a significant decline in Food and a trending decline in Consumables. Conversely, it demonstrates a significant lift in the ETS and, more generally, positive changes across most General Merchandise verticals. This corroborates that short-term behavioral features are more informative in an environment that is more dynamic in terms of both inventory assortment and customers' shopping behaviors. In contrast, long-term features are more advantageous for business units Food and Consumables, which typically display more stable and enduring shopping patterns. Therefore, it is very tempting to employ both types of features in the ranking model to leverage their combined strengths. Similar ideas of combining session and historical customer search behaviors per each customer are also investigated in web search personalization [20], but to the best of knowledge, our work is the first one to explore combining (query, product)-level historical behavior features over different lookback time windows in product search ranking.

2.2. Both Long & Short Windows

With the insight from the previous subsection, we set up the second interleaving test as follows:

- **Control:** Baseline Model (2-year behavioral features only),
- **Variant:** Model B (2-year and 1-month behavioral features),

with the purpose to examine the combined impact of using both 2-year and 1-month behavioral features in ranking.

The test result is presented in Table 3. To our surprise, Model B performs quite sub-optimally overall with the degradation in Food, Consumables, and ETS verticals. This suggests that combining both long- and short-term behavioral features in this vanilla manner not only fails to provide gains in ranking performance but also leads to further declines. One possible reason for the negativity is the lack of flexibility in our ranking model to leverage different behavioral features accordingly. For instance, the Food vertical should ideally leverage the 2-year behavioral features as extensively as possible. However, adding 1-month features dilutes the positive effect of the 2-year features, negatively interfering with the overall contribution of behavioral features. Conversely, in verticals such as ETS and Hardlines, where 1-month features are more advantageous, the inclusion of 2-year features can similarly impair performance.

Vertical	Food	Consumables	Home	ETS	Hardlines	Fashion	Overall
Change	−0.46%*	−0.51%*	+0.29%	−0.46%	−1.07%	+1.72%	−0.41%*

Table 3
Result of % searches with engagement for Test 2. The control is the baseline model (only using 2-year behavioral features), and the variant is Model B (using both 1-month and 2-year behavioral features).

3. How to Integrate Both?

Different verticals exhibit different patterns of trending effects in customer behaviors. For instance, Fashion such as "clothes" may be significantly influenced by recent trends affecting their popularity and customer interactions. In contrast, Food and Consumables such as "milk" and "toilet paper" tend to show more stability over time and are predominantly shaped by long-term engagement patterns.

Based on observations from tests in Sections 2.1 and 2.2, to improve the model performance with combined behavioral features of both long and short windows, we consider making our ranking model more query context-aware by incorporating one-hot encoded query vertical signals (predicted from the upstream query understanding model) into the model. These query-level vertical signals would better guide our ranking model to leverage behavioral features of different time windows according to different queries. Thus, we propose the fourth ranking model below.

- **Model C:** model with both 2-year and 1-month features, and query-level vertical features.

3.1. Both Long & Short Windows with Verticals

The third interleaving test is configured as follows.

- **Control:** Baseline Model (2-year behavioral features only),
- **Variant:** Model C (2-year and 1-month behavioral features with the vertical features),

with the purpose to examine whether adding query-level vertical features helps better integrate 2-year and 1-month behavioral features in ranking.

The test result, detailed in Table 4, shows that guided by vertical information, behavioral features are more effectively utilized by the ranking model, leading to significant uplifts across all General Merchandise verticals while rectifying the previous degradation in the Food and Consumables. Model C also shows an overall significant increase of 0.22% in customer engagement and proves to be the best candidate ranking model among all tested. This demonstrates that incorporating vertical features can indeed enhance the integration of multi-window behavioral features, allowing each to play to its strengths and mitigate its weaknesses.

3.2. Why Does Long & Short & Verticals Work?

The guiding effect that vertical information has on the ranking model in using different behavioral features can also be validated in the model structure. Our ranking model inherently

Vertical	Food	Consumables	Home	ETS	Hardlines	Fashion	**Overall**
Change	-0.01%	$+0.02\%$	$+0.40\%^*$	$+0.78\%^*$	$+1.58\%^*$	$+0.73\%^*$	$+0.22\%^*$

Table 4
Result of % searches with engagement for Test 3. The control is the baseline model (only using 2-year behavioral features), and the variant is Model C (using both 1-month and 2-year behavioral features along with the vertical features).
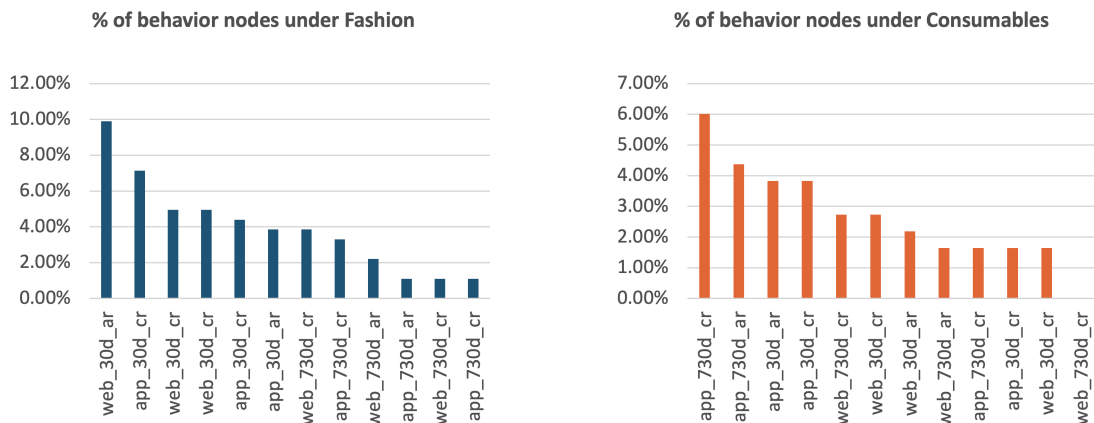

Figure 1: Distribution of different behavioral nodes under vertical nodes for Fashion and Consumables. We summarize the percentage of behavioral tree nodes under the verticals. The behavioral feature names have 3 segmented sub-strings: the first indicates the behavioral data source (web and app); the second indicates the time window lengths (730-day and 30-day); the third indicates the behavioral types (cr = click rate; ar = ATC rate; or = order rate).

employs a tree structure, where adjacent tree nodes tend to be functionally related. More precisely, if the nodes corresponding to one feature frequently precede those of another specific feature, it suggests that the former, i.e., the upper-level feature, exerts a certain degree of influence/control over the latter, i.e., the lower-level feature, determining when it will activate to impact the model's predictions.

Across all splitting nodes from the trees in Model C, we summarize the distribution of different behavioral nodes under the vertical nodes in Figure 1, taking Fashion and Consumables as examples. The results show that 1-month behavioral features are more influential for Fashion queries since they more prevalently occupy the place of the immediate lower level when the current vertical node is Fashion, whereas 2-year behavioral features are more prevalent for Consumables queries. This observation is aligned with our interpretation of the test result in Section 3.1 given the characteristics of different verticals, and it evinces that introducing query-level vertical signals can help guide our ranking model to better ensemble long- and short-term behavioral features in the sense that different behavioral features can contribute accordingly with respect to different search queries.

Metric	Overall GMV	Marketplace GMV	ATC@10	Sessions with ATC	Session Abandonment
Lift	+0.12%	+0.64%*	+0.21%*	+0.22%*	−0.16%*

Table 5

Online A/B test result of Model C vs. baseline. These metrics are all calculated at visitor level: 1) **Overall GMV**: the total Gross Merchandise Value from all kinds of products sold; 2) **Marketplace GMV**: the Gross Merchandise Value yielded from marketplace products; 3) **ATC@10**: the percentage of search ATCs coming from top 10 products in search results; 4) **Sessions with ATC**: percentage search sessions with at least one ATC; 5) **Session Abandonment Rate**: percentage search sessions without any user engagement (and thus the smaller the better).

4. A/B Test

After the series of interleaving tests in Sections 2 and 3, we decided to move forward to A/B test with the most promising candidate, Model C, which incorporates both long- and short-term behavioral features along with the query-level vertical signals. Specifically, we conducted a comprehensive A/B test on *Walmart.com* for two weeks to compare Model C against the baseline production model.

The result, detailed in Table 5, highlights substantial improvements in key search related metrics. This A/B test observation confirms our hypothesis that a vertical-aware ranking model incorporating a hybrid of behavioral features across both long and short time windows can enhance the customer experience for a diverse range of online shopping needs. In addition, the positivity in marketplace GMV clearly indicates that we are also able to better address cold-start problems when introducing short-term behavioral features into the system.

We also present a qualitative example in Figure 2 illustrating the comparison of Model C versus the baseline in terms of user experience from the search ranking. It is clearly demonstrated that utilizing behavioral features from both long and short time windows, along with vertical information, results in a ranking model that prioritizes products with high recent popularity, especially in the General Merchandise categories. This approach ensures that customers are provided with options that are more closely aligned with their current shopping needs.

5. Conclusion

In this paper, we propose a novel product search ranking model that incorporates a hybrid of behavioral features over both long and short lookback time windows with vertical-specific insights. The multi-window design aims to capture customer engagement patterns over varying durations, and the vertical features are purposed to tailor behavioral features more effectively to different online shopping contexts. This approach allows long-term behavioral features to reflect enduring patterns, supporting routine customer journeys, while short-term features capture immediate, trending patterns to enhance discovery customer experiences.

Through comprehensive online testing, we demonstrate that the proposed model significantly outperforms the baseline, which solely utilizes singular time-window behavioral features, by achieving substantial improvements in key evaluation metrics across various verticals, catering to distinct online shopping needs. As a result, the integration of multi-window behavioral
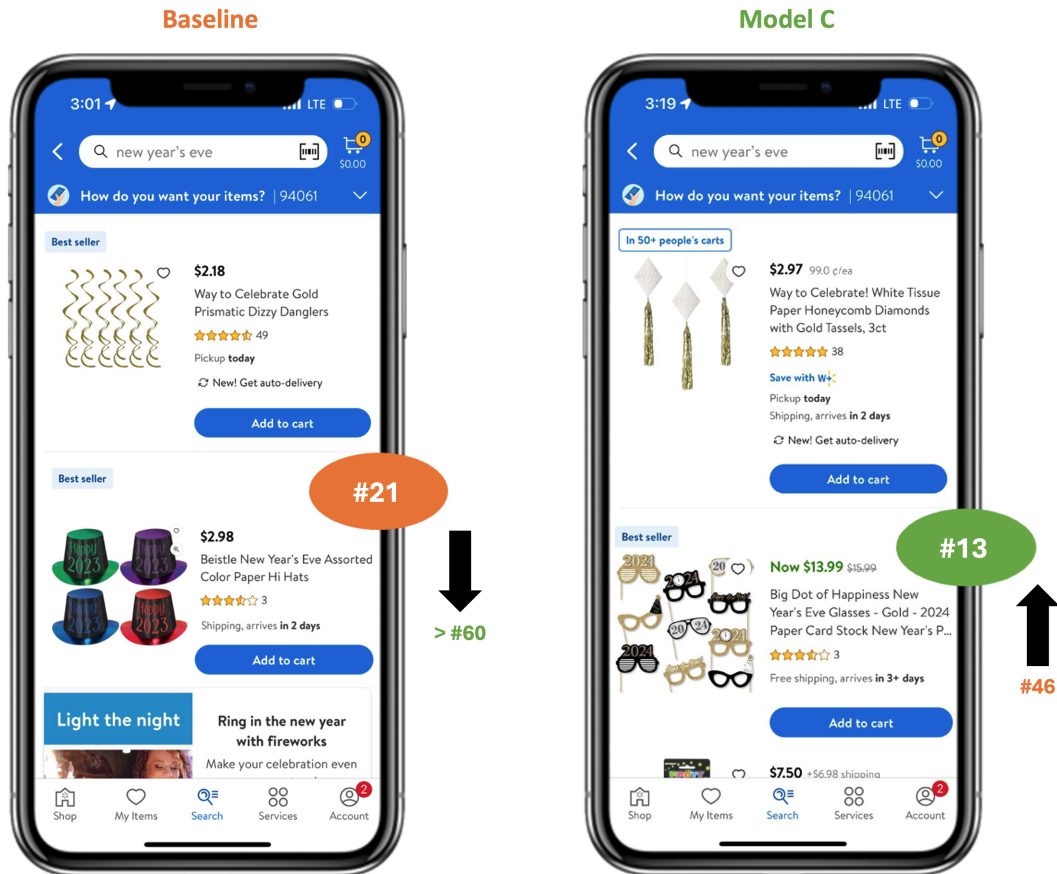
Figure 2: Example of customer search ranking experience comparing baseline vs. Model C.
This comparison was performed in December 2023 for query *new year's eve*. The baseline incorrectly prioritizes many 2023 New Year's Eve supplies over 2024 products, which should have been ranked higher as the event for 2024 approaches. In contrast, Model C effectively ranks 2024 products higher than those for 2023. In this example, the 2024 glasses item is elevated from position 46 to 13, while the 2023 hat item is demoted from position 21 to beyond 60, moving it off the first page of search results.

features and search context awareness adeptly navigates the complex dynamics of different shopping categories, thereby enhancing customer engagement across all verticals. Consequently, the proposed model not only fulfills the diverse needs of contemporary eCommerce online shopping but also lays a scalable foundation for future enhancements in search ranking systems.

For future work, we intend to expand the feature scope of the search ranking model by incorporating behavioral features from additional time windows, such as 1 week and 1 year. This extension will enable the model to capture a broader spectrum of trending effects, further enhancing its predictive accuracy. Additionally, we plan to introduce more granular query-level signals–e.g., categorical signals, product type signals, etc.–to allow for more nuanced guidance of behavioral features, improving ranking's contextualized capability and enriching the online shopping experience for customers.

References

[1] D. Sorokina, E. Cantu-Paz, Amazon search: The joy of ranking products, in: SIGIR, 2016, pp. 459–460.

[2] A. Trotman, J. Degenhardt, S. Kallumadi, The architecture of ebay search, in: SIGIR eCom, volume 2311, 2017.

[3] E. P. Brenner, J. Zhao, A. Kutiyanawala, Z. Yan, End-to-end neural ranking for ecommerce product search, in: SIGIR eCom, 2018.

[4] M. Tsagkias, T. H. King, S. Kallumadi, V. Murdock, M. Rijke, Challenges and research opportunities in ecommerce search and recommendations, in: ACM SIGIR Forum, 2021.

[5] R. Eletreby, C. Mu, Z. Wang, R. Mukherjee, Machine learning based methods and apparatus for automatically generating item rankings, 2022. US Patent App. 17/246,179.

[6] X. Wu, A. Magnani, S. Chaidaroon, A. Puthenputhussery, C. Liao, Y. Fang, A multi-task learning framework for product ranking with bert, in: WWW, 2022, pp. 493–501.

[7] C. Burges, R. Ragno, Q. Le, Learning to rank with nonsmooth cost functions, NeurIPS (2006).

[8] C. Burges, From ranknet to lambdarank to lambdamart: an overview, Learning 11 (2010) 81.

[9] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

[10] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, X. Cheng, A deep look into neural ranking models for information retrieval, Information Processing & Management 57 (2020) 102067.

[11] O. Chapelle, Y. Chang, Yahoo! learning to rank challenge overview, in: PMLR, 2011, pp. 1–24.

[12] P. Gupta, T. Dreossi, J. Bakus, Y. Lin, V. Salaka, Treating cold start in product search by priors, in: WWW Companion, 2020.

[13] C. Han, P. Castells, P. Gupta, X. Xu, V. Salaka, Addressing cold start in product search via empirical bayes, in: CIKM, 2022.

[14] M. Hendriksen, E. Kuiper, P. Nauts, S. Schelter, M. de Rijke, Analyzing and predicting purchase intent in e-commerce: Anonymous vs. identified customers, in: SIGIR eCom, 2020.

[15] J. J. Rocchio, Relevance feedback in information retrieval, The SMART retrieval system: experiments in automatic document processing (1971).

[16] K. Santu, S. Kanti, S. Parikshit, C. Zhai, On application of learning to rank for e-commerce search, in: SIGIR, 2017, pp. 475–484.

[17] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: KDD, 2016, pp. 785–794.

[18] T. Y. Liu, Learning to rank for information retrieval, Foundations and Trends® in Information Retrieval 3 (2009) 225–331.

[19] O. Chapelle, T. Joachims, F. Radlinski, Y. Yue, Large-scale validation and analysis of interleaved search evaluation, ACM Transactions on Information Systems 30 (2012) 1–41.

[20] P. Bennett, R. White, W. Chu, S. Dumais, P. Bailey, F. Borisyuk, X. Cui, Modeling the impact of short- and long-term behavior on search personalization, in: SIGIR, 2012, pp. 185–194.