Identifying High Consideration E-Commerce Search Queries

Zhiyu Chen, Jason Choi, Besnik Fetahu, Shervin Malmasi

Amazon.com, Inc., Seattle, WA, USA

{zhiyuche,chojson,besnikf,malmasi}@amazon.com

amazon | science

Introduction

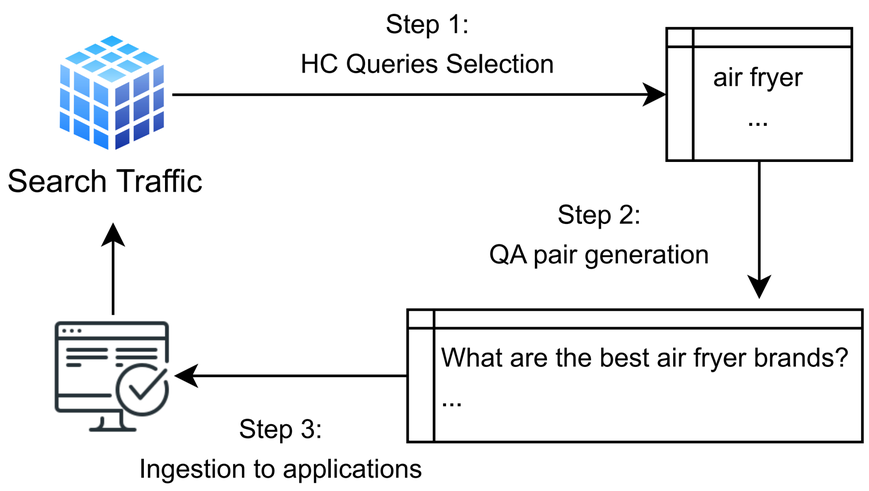
- The effective identification of specific keywords is essential not only for driving organic traffic and revenue in e-commerce, but also for enhancing the overall customer experience.

➢ Example1 : when customers search for "prime day deals" on Amazon, a dedicated widget will appear above product search results.

➢ Example 2: a Question-Answer (QA) widget with relevant shopping knowledge could be rendered to match the customer's query.

An example of a QA component in search results for query "air fryer".

- Identifying such queries is usually the initial step prior to content creation and targeting.



- We refer to such searches as **High-Consideration (HC) Queries** since consumers require additional information to consider their decision, or refine their search.

- We explore the task by proposing an **Engagement based Query Ranking (EQR)** approach.

Method

- We model the task of HC queries identification as a ranking task.

- We first define our proxy target, the query (q) level engagement score (e), as following:

$$e(q) = \frac{freq_c(q)}{freq(q)} \qquad (1)$$

- $freq_c(q)$ is the user click count for an informational component (e.g. QA widget) displayed in the search results.
- $freq(q)$ is the total frequency for the query.

- Note that the definition of the engagement score can vary among different businesses.

- We consider the overall query-level engagement instead of content-level engagement like the CTR of an individual QA pair generated for a query.

- For a query $q \in Q$, we have its user interaction features $x \in R^d$ where d is the total number of features.

- We aim to learn a function f(·) that could predict the engagement score of a query given its features. Once we obtain the predicted engagement scores for a list of n queries $Q = \{q_1, \ldots, q_i, \ldots, q_n\}$, then we re-rank them in descending order of their engagement scores.

Query Features

- **Behavioral Features** characterize user interactions with a search system for a query. After a query is submitted, we observe subsequent interactions such as clicking on a search result, or adding a product to the cart.

- **Financial Features** relate to the purchases associated with a query. We hypothesize that financial signals (e.g. order volumes, prices, temporal patterns) can help distinguish HC queries.

- **Catalog Features** focus on features of the products served in the search results, as derived from the product catalog. Such features serve as feedback from the product search system and are inspired by post-retrieval methods for predicting query performance.

- See detailed feature descriptions in the paper.

Training

- We train Gradient Boosted Decision Trees (GBDT) to predict query engagement scores:

$$\mathcal{L} = \frac{1}{2}\sum_i (f(x_i) - e(q_i))^2 \qquad (2)$$

- We collected the data associated with a QA component from Amazon spanning a one-year period.

- We obtained the corresponding query-level user engagement data as our proxy targets. The data was divided into training, validation, and test sets, with respective proportions of 70%, 15%, and 15%

Evaluation

- Decision Tree Ensembles yield best overall performance.

- Text-based methods underperform feature-based models. Even GPT-3.5 performs poorly in all metrics, with little or no improvement even with few-shot in-context learning.

- Behavioral features are of the most importance, followed by financial features.

Method	Hit@5	Hit@50	Hit@100	Hit@500	Kendall's Tau	MSE
Frequency	0.00	0.04	0.05	0.17	0.34	-
XGBoost (all features)	**0.20**	**0.42**	**0.50**	**0.69**	**0.52**	**0.0038**
XGBoost (behavioral only)	0.00	0.26	0.39	0.63	0.50	0.0041
XGBoost (financial only)	0.00	0.18	0.17	0.55	0.43	0.0049
XGBoost (catalog only)	0.00	0.08	0.14	0.46	0.17	0.0063
Random Forest	**0.20**	0.36	0.37	0.67	0.51	0.0040
Lasso	0.00	0.34	0.34	0.62	0.46	0.0043
Ridge	0.00	0.30	0.32	0.60	0.47	0.0044
Elastic Net	0.00	0.34	0.32	0.63	0.47	0.0043
Linear	0.00	0.32	0.31	0.61	0.47	0.0043
RoBERTa	**0.20**	0.28	0.34	0.50	0.32	0.0112
GPT-3.5 (zero-shot)	0.00	0.06	0.11	0.30	0.05	0.4024
GPT-3.5 (few-shot)	0.00	0.06	0.08	0.30	0.05	0.1049
GPT-4o (zero-shot)	0.00	0.16	0.18	0.40	0.14	0.3822
GPT-4o (few-shot)	0.00	0.14	0.14	0.41	0.12	0.0486

- Human evaluation: the precision of our model is 96%.

- Online deployment: the model-chosen queries outperformed the human-selected set with a relative increase of 6%.

- High Consideration Queries: iphone pro max, nikon d500, dji mini
- Low Consideration Queries: rubber mats for gym, pink belt, purse strap

Conclusions

- We introduce the task of Engagement-based Query Ranking in order to select High Consideration queries.

- We proposed three categories of features to train pointwise rankers to address this task.

- Our experimental results show that our proposed method achieves better performance than the baselines.

- The human evaluation indicates that our method could serve as an effective tool to save resources spent on error-prone human annotations.