Why Do People Buy Seemingly Irrelevant Items in Voice Product Search?
On the Relation between Product Relevance and Customer Satisfaction in eCommerce

David Carmel
Amazon
Haifa, Israel
dacarmel@amazon.com

Elad Haramaty
Amazon
Haifa, Israel
eladh@amazon.com

Arnon Lazerson
Amazon
Haifa, Israel
arnonl@amazon.com

Liane Lewin-Eytan
Amazon
Haifa, Israel
lliane@amazon.com

Yoelle Maarek
Amazon
Haifa, Israel
yoelle@yahoo.com

ABSTRACT
One emerging benefit of voice assistants is to facilitate product search experience, allowing users to express orally which products they seek, and taking actions on retrieved results such as adding them to their cart or sending the product details to their mobile phone for further examination. Looking at users’ behavior in product search, supported by a digital voice assistant, we have observed an interesting phenomenon where users purchase or engage with search results that are objectively judged irrelevant to their queries.

In this work, we analyze and characterize this phenomenon. We provide several hypotheses as to the reasons behind it, including users’ personalized preferences, the product’s popularity, the product’s indirect relation with the query, the user’s tolerance level, the query intent, and the product price. We address each hypothesis by conducting thorough data analyses and offer some insights with respect to users’ purchase and engagement behavior with seemingly irrelevant results. We conclude with a discussion on how this analysis can be used to improve voice product search services.

CSC CONCEPTS
• Information systems → Search interfaces.

KEYWORDS
product search, customer engagement, objective relevance judgment

1 INTRODUCTION
Search engines invest a lot of efforts in retrieving high quality results for their customers. The quality of results is typically measured by their relevance to the query. We consider the domain of voice product search, supported by digital voice-based assistants, and focus on user engagement with the search results. In this domain, user’s needs are both informational (when they ask questions about a specific product or a family of products) or transactional (when they want to acquire a specific product). The fact that in voice, both the input and output are spoken, results in customers being exposed to fewer results with much less information [8]. This is even more pronounced in devices without a screen, called “headless devices”, which generate a growing portion of the voice traffic. The customer’s attention in such voice interactions will typically not go beyond one or two candidates, which imposes higher quality constraints on the top results than in the traditional Web search scenario.

While we would naturally expect that a relevant product be what users seek, we have examined the query logs of a voice assistant supporting shopping needs and encountered a surprising phenomenon illustrated by the following example. Table 1 shows two different products offered by our voice assistant for the query “buy burger”. Interestingly, the relevant product of “Angus Burgers” was never purchased, while the irrelevant “Burger Press” was purchased by several customers.

Indeed, we have observed that in a non-negligible number of cases, users who searched for products on headless devices significantly engage with irrelevant search results. The term “irrelevant” may be misleading here since a relevant item is typically interpreted as “anything that satisfies the user’s needs”. Thus, the title of this work may look as an oxymoron – the purchase of a product is a strong signal of user’s need and should therefore be an evidence of relevance. In the context of this work we mark product items as relevant or irrelevant to the user’s query based on objective relevance judgment of human annotators.

The natural question addressing why irrelevant items were retrieved by the search engine in the first place is out of the scope of this work. However, we can safely assume that most eCommerce
Table 1: A pair of products offered for the query ‘buy burger’. The seemingly irrelevant item (burger press) has a higher purchase level.

<table>
<thead>
<tr>
<th>Product</th>
<th>Relevant?</th>
<th>Purchase Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 (6 oz.) Angus Burgers</td>
<td>Yes</td>
<td>0.0</td>
</tr>
<tr>
<td>Stuffed Burger Press</td>
<td>No</td>
<td>0.03</td>
</tr>
</tbody>
</table>

search engines are trained to optimize for user’s engagement and conversion, possibly at the cost of relevance in some cases. Moreover, ambiguous queries, or errors in query transcription, may also be followed with irrelevant search results.

Extensive research was conducted in the domain of product search, from the perspective of the search engine, to improve search quality while reducing the amount of irrelevant search results (e.g., [1, 3, 4, 9, 19, 21]). We examine here on the problem from the user’s perspective. We analyze the phenomenon of user engagement with seemingly irrelevant search results and demonstrate its scope and significance. In the following, we address several hypotheses as to the reasons behind it:

- **Personalized preferences.** Relevance is a subjective notion, thus, an (objectively judged) irrelevant product might be considered relevant by the customer.
- **Popularity.** The product popularity is a significant factor in a purchase decision. This might affect customers’ purchase or engagement with irrelevant popular results.
- **Exploration mode.** Customers do not always seek to buy products [2]. Sometime they explore the catalog to get some shopping ideas. We can expect high engagement with irrelevant results when customers are in exploration mode.
- **Tolerance.** There are several circumstances in which customers are more tolerant of irrelevant results. For example, when queries are unspecified, e.g., “buy coffee”, or when queries are over-specified, e.g., “buy home basic sink collapsible colander one size gray and white”. In such cases when exact results are not well defined, or when there is no product in the catalog that answers all specified constraints, the customer may be satisfied with the engine’s offer.
- **Price.** Product price is a significant factor in customers’ purchase decision. We hypothesize that the more expensive the irrelevant product is, the lower its purchase likelihood, and vice versa. For example, as evident in our query log. searching for a golden Rolex watch (~$100,000), mostly ends with a purchase of a much cheaper watch.
- **Indirect Relation** Several factors may lead to customers’ perception that a seemingly irrelevant product is related to their query, through sharing some characteristics with the query’s objectively relevant products. They could be of the same type, belong to the same category, or often be purchased together. We hypothesize that the more “similar” a product to the objectively relevant products of the query, the higher the engagement level will be.

Table 2 presents additional anecdotal examples from our query log where the purchased product was judged irrelevant to the user’s query. For each query-product pair, we predict the reason behind its purchase. We note that, in most cases, the reason may be a combination of several hypotheses, but, to avoid clutter, we present only one for each case. In the rest of the paper we conduct data analyses to verify each hypothesis and present the insights we derived from them. We conclude with a discussion on how this kind of analysis can be used to improve a voice product search service.

### 2 RELATED WORK

Recent research on voice-based search [7] has been conducted in the context of mobile devices, showing that users issue much longer queries, and the language of voice queries is closer to natural language than typed queries. However, to the best of our knowledge, no previous research has focused on user behavior analysis in the context of product search served by voice assistants.

Product search provided by e-commerce sites has also attracted a lot of attention in recent years. Many studies investigated the relations between relevance and purchase likelihood [2, 10, 11, 13, 17, 21]. Alonso and Mizzaro [2] presented evidence that when customers are interested in buying products, they apply many criteria in addition to relevance. Kumar et al. [10] observed that predicting the query performance in product search cannot be based only on standard metrics such as click-through-rate (CTR), since CTR might be high while the results are poor, in terms of relevance perspective. Su et al. [17] observed that the results expected by two different users for the same query may be different and thus they may be individually dissatisfied even though the results seem to be relevant. They presented different user interaction patterns and demonstrated that user satisfaction can be predicted by utilizing the interaction behavior.

Sondhi et al. [16] suggested a taxonomy of product queries by analyzing an e-commerce search log. They showed that each category can be associated with distinctive user search behavior. Guo et al. [6] proposed an attentive Long Short-Term Preference model for personalized product search, by integrating long- and short-term user preferences with the current query. Two attention networks were designed to distinguish short-term factors as well as long-term user preferences, thus capturing users search intentions more accurately.

In this work we consider two different types of affirmative behaviors, performed by the customers, over the search results provided by a voice digital assistant in the product search domain. One is purchase, a strong signal of user satisfaction. The second is engagement which includes all other engagement actions such as add-to-cart or send-to-phone. We offer some insights that will hopefully provide a better understanding of user behavior in the voice shopping domain.
Why Do People Buy Seemingly Irrelevant Items in Voice Product Search?  

<table>
<thead>
<tr>
<th>“buy..”</th>
<th>Product</th>
<th>Hypothesis</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘coffee’</td>
<td>Dixie Paper cups and lids, 156 count of 6 pack</td>
<td>Personalized preferences</td>
<td>Item was returned from user order history</td>
</tr>
<tr>
<td>‘pizza’</td>
<td>Gummi Pizza by E-Fruitti 48 Count (Net Wt. 26oz)</td>
<td>Trendiness &amp; Popularity</td>
<td>Frequently bought item for the query</td>
</tr>
<tr>
<td>‘video games’</td>
<td>High-Speed HDMI Cable, 6 Feet, 1-Pack</td>
<td>Exploration mode</td>
<td>Query is broad</td>
</tr>
<tr>
<td>‘zone perfect nutrition snack bars strawberry yogurt 36 count’</td>
<td>ZonePerfect Strawberry Yogurt Nutrition Bars, 30 count of 1.76 oz each</td>
<td>Tolerance</td>
<td>Tolerance to quantity difference</td>
</tr>
<tr>
<td>‘iphone 10’</td>
<td>Apple iPhone 6, GSM Unlocked, 16GB - Gold (Renewed)</td>
<td>Price</td>
<td>Similar product type as the iPhone 10, but cheaper</td>
</tr>
<tr>
<td>‘apple watch’</td>
<td>Apple Watch Stand, iPhone Stand, BENTO-BEN Iwatch Charging Stand Dock Station Cradle - Rose Gold</td>
<td>Relatedness</td>
<td>Highly related accessory</td>
</tr>
</tbody>
</table>

Table 2: Seemingly irrelevant products purchased by customers.

3 PRELIMINARIES

We consider \((q, p)\) pairs, where \(p\) is the product offered by our system in response to the customer’s query \(q\). For each such pair, we measure the relevance of \(p\) to \(q\), and note the action that the customer performed in response to offer \(p\). These measurements are described in more details below.

Relevance. What makes a search result relevant (or irrelevant) to a search query? It is well accepted in the IR research community that annotators tend to agree on the relevance of a document to a query, to some extent [20]. In the context of this work, we identify the relevance of a product \(p\) to the customer’s query \(q\) based on the relevance judgments of at least three human annotators. The annotators were directed to judge the relevance of \(p\) to \(q\) by estimating their shopping satisfaction in a hypothetical scenario where they ask a shopping agent to shop for \(q\), and get \(p\) in response. Relevance is determined based on the majority vote. We mark \(R(q, p)\) in case \(p\) is judged relevant to \(q\). We mark \(R(S) = \{(q, p) \in S | R(q, p)\}\) be the subset of all relevant pairs in the set \(S\).

Customer actions. For each \((q, p)\) pair, we denote the action that the customer performed in response by \(a((q, p))\), where \(a((q, p)) \in \{\text{ignore}, \text{add2cart}, \text{send2phone}, \text{purchase}\}\); ignore indicates the user ignored offer \(p\), purchase indicates a purchase action, add2cart and send2phone indicate engagement actions. For simplification, we do not distinguish between engagement actions and consider them as having equal importance.

We define the purchase level of a pairs set by:

\[
PL(S) = \frac{\#\{(q, p) \in S | a((q, p)) = \text{purchase}\}}{|S|}.
\]  

(1)

and the engagement level of a pairs set by:

\[
EL(S) = \frac{\#\{(q, p) \in S | a((q, p)) \in \{\text{add2cart, send2phone}\}\}}{|S|}.
\]  

(2)

Given a pair set \(S\), we distinguish between the purchase level and the engagement level for the relevant pairs \(R(S)\), and for the complementary subset of irrelevant pairs, \(IR(S)\). Let \(pRatio(S) = \frac{PL(IR(S))/PL(R(S))}{|S|}\) be the purchase ratio between the purchase level of the irrelevant pairs set and the purchase level of the relevant set. Similarly, \(eRatio(S) = EL(IR(S))/EL(R(S))\) be the engagement ratio between the two subsets. The higher the ratio is, the more people purchase (or engage) with irrelevant products, compared to relevant ones. For example, when the ratio is 0, people purchase (or engage) only with relevant product, while when it is 1, they are indifferent to the relevance of the product to the query.

Finally, we define the normalized purchase level \(NPL(S)\), and the normalized engagement level \(NEL(S)\), by normalizing the purchase/engagement level with the overall purchase/engagement level of all the relevant pairs:

\[
NPL(S) = \frac{PL(S)}{PL(R(All))}, \quad NEL(S) = \frac{EL(S)}{EL(R(All))}.
\]  

(3)

In the following, we analyze the difference in purchase and engagement levels for different sets of pairs, and propose some explanations for the reasons behind it.
Table 3: Normalized purchase and engagement levels for irrelevant pairs (IR), for different product sources and query intents.

<table>
<thead>
<tr>
<th>Subset (S)</th>
<th>NPL(S)</th>
<th>NPL(IR(S))</th>
<th>pRatio</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1.00</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>Personalized Source</td>
<td>2.54</td>
<td>0.42</td>
<td>0.16</td>
</tr>
<tr>
<td>Popularity-based Source</td>
<td>0.29</td>
<td>0.08</td>
<td>0.28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subset (S)</th>
<th>eRatio</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1.0</td>
</tr>
<tr>
<td>Personalized Source</td>
<td>0.9</td>
</tr>
<tr>
<td>Popularity-based Source</td>
<td>1.1</td>
</tr>
</tbody>
</table>

(a) Purchase levels

Table 3: Normalized purchase and engagement levels for subsets of (q,p) pairs, split between relevant pairs (R(S)) and irrelevant pairs (IR(S)), for different product sources and query intents.

<table>
<thead>
<tr>
<th>Subset (S)</th>
<th>NEL(S)</th>
<th>NEL(IR(S))</th>
<th>eRatio</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1.0</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Personalized Source</td>
<td>0.9</td>
<td>1.0</td>
<td>1.1</td>
</tr>
<tr>
<td>Popularity-based Source</td>
<td>1.1</td>
<td>0.7</td>
<td>0.63</td>
</tr>
</tbody>
</table>

(b) Engagement levels

We also note that there is a difference in purchase and engagement levels among sources. For the personalized source, both NPL(S) and NPL(IR(S)) are extremely high compared to other sets, showing that previously ordered products are more likely to be re-purchased by our customers, even if irrelevant to the query. Furthermore, engagement with irrelevant pairs of the personalized source is larger than with those from the Popularity-based source. The more personalized the search results are, the higher the engagement with irrelevant results, due to the fact that an objective relevance judgment is more likely to deviate from the customer’s personalized relevance judgment. For the popularity-based source, while the purchase level is significantly low compared to the personalized source, we observe that the pRatio is significantly higher, showing that customers purchase (relatively) more irrelevant results, with respect to relevant ones, indicating that product popularity is another significant factor affecting purchase decision.

4.2 Exploration Intent

We analyze the subset of queries reflecting a buy intent (e.g., “buy iphone 8 case”) and queries reflecting an exploration intent (e.g. “find iphone 8 case”). The query intent was determined using an external classifier used by our search engine.

Like in Section 4.1, each of these subsets contains more than 1M (q,p) pairs, and was further split to relevant (R(S)) and irrelevant (IR(S)) pairs. Table 3 presents the purchase level and the engagement level, for the partition of the whole set of pairs (“All”), as well as for the two intents, Buy and Exploration. Results were normalized with respect to the purchase/engagement level of R(All). It can be seen that the normalized purchase level of Exploration intent is much lower than the normalized purchase level of Buy intent (which reflects an explicit purchase need), both for relevant and irrelevant pairs. The pRatio and eRatio for Exploration intent are higher than for Buy intent, showing that in exploration mode, customers purchase (relatively) more irrelevant results, and are more engaged with irrelevant results. However, unexpectedly, there is no significant difference in engagement level between relevant and irrelevant pairs for queries having different intents.

4.3 Tolerance

Following our assumption that the customer tolerance to irrelevant search results varies with query specificity, we measured the pRatio as a function of the query length\(^2\) as presented in Figure 1. A similar trend was obtained for eRatio hence is not presented due to lack of space.

It can be seen that pRatio is high for one-term queries, as well as for long queries with five or more terms. As expected, for very short queries, often not well specified, people are more tolerant with the search results, as reflected by the high pRatio. This is also true for over-specified queries where the query is long. Looking at a small sample of verbose queries, we have noticed that in many cases people tend to provide too many details (e.g. many product attributes) that are not necessarily needed for identifying the relevant product. This might partially explain the fact that customers are more tolerant with irrelevant results that do not cover all the attributes specified in these queries.

\[^2\]Following our strict privacy preserving mechanisms, only frequent queries, issued by at least 10 unique users, were annotated and retained in our sample. This added a bias to our dataset toward head queries. Purchase/engagement actions that were later canceled by the customers were not counted to reduce the effect of unintentional actions.

\[^3\]The query length was measured by the actual number of terms submitted to the search engine, i.e. after stop-word filtering and keyword extraction.
Another factor related to customers’ tolerance of irrelevant search results is the product category. Indeed, it turns out that tolerance of irrelevant results varies between different categories. We measured the pRatio and eRatio across several product categories, presented in Figure 2. It can be seen that there is a high variance in the ratios across categories. In addition, it can be seen that there is low correlation between purchase ratio and engagement ratio across the categories. Some categories (toys, digital products, pet products, drugstore) have an engagement ratio larger than one, meaning that people are more engaged with irrelevant products than with relevant ones. In particular, the ratio is extremely large for toys and digital products, probably due to the entertainment nature of such products. In contrast, it seems that in some categories such as beauty and groceries, people expect to get the exact product they ask for and are intolerant with irrelevant results.

4.4 Product Price

In order to estimate the effect of the product price on purchase level, we compared the price distribution of purchased items in the relevance subset and in the irrelevance subset of all (q, p) pairs.

Let \( P(q) \) be the set of products purchased for a query \( q \), and \( price(p) \) the price of product \( p \). The average purchase price for query \( q \) is \( AP(q) = \frac{1}{|P(q)|} \sum_{p \in P(q)} price(p) \). For each (q, p) pair, we measured its normalized purchase price, \( npp(q, p) = \frac{price(p)}{AP(q)} \).

Figure 3 shows the histogram of the number of pairs, split into bins of equal sizes according to their \( npp \) value.

While the price distribution of relevant pairs is close to the normal distribution over the average \( npp \) price (around 0.8), the price distribution of irrelevant pairs is strongly skewed to the left (around 0.6), meaning that the purchase level for irrelevant items is biased toward low-price offers. In general, a purchase decision for low-price products is easier for customers, in particular for irrelevant products. Figure 4 shows a similar behavior for engagement price distribution; with a normal price distribution (around 0.8) for relevant pairs, and a strongly skewed to the left distribution (around 0.4) for irrelevant pairs. Consequently, as could be expected, the lower the price of an irrelevant product, the higher the likelihood of purchasing or engaging with it.

4.5 Indirect relation

The objective relevance judgment used in this work captures only a narrow aspect of product relevance. There are many other factors that cause users to associate a product with a query. For example, a seemingly irrelevant product might have an “indirect” relation with the query, namely, sharing some characteristics with products that are objectively relevant to the query. Examples of such characteristics include sharing the same product type, brand, or category, or being frequently purchased together. Our hypothesis is that the purchase likelihood of a seemingly irrelevant product increases with its level of similarity to the query’s relevant products.

In order to assess the level of indirect relation between a query and a product, we use two measures of similarity between products. The first one, Semantic Similarity, estimates the semantic similarity between the product descriptions, which provides a proxy for the relation between the products’ type, brand, and category. The second one, Purchase Similarity, measures the likelihood of different products being purchased by ‘similar’ users. We next elaborate on these two similarity measures. For both measurements we define the indirect relation between a product and a query based on the product similarity to the query’s (objectively judged) relevant products.
Semantic Similarity: Let desc(p) be the textual product description of product p which is a concatenation of the product title with the product type. The semantic similarity between two products p1, and p2, is defined by

SemSim(p1, p2) = \cos(w2v(desc(p1)), w2v(desc(p2))),

where w2v(t) is the centroid of the word embedding representations of all words in text t, and \cos(\cdot) is the cosine similarity between vectors. For word embedding we used the pre-trained Glove embedding model\(^4\). An example for a pair of two semantic similar products is the 1998 thriller movie “Twilight” and the 2008 fantasy movie “Twilight”.

Let R(q) be the set of all relevant products to the query q. We define the description of query q to be the concatenation of descriptions of its relevant products desc(q) = \sum_{p \in R(q)} desc(p). The indirect relation between query q and product p is defined by SemRel(q, p) = \cos(w2v(desc(q)), w2v(desc(p))).

Purchase Similarity: We consider two products share some purchase similarity if they tend to be bought by similar users, while similar users are defined as users who tend to buy similar products. This cyclic definition is similar in nature to semantic word similarity in which, words are similar if they tend to appear in proximity to similar words [14].

More formally, products are embedded into a low-dimensional space based on historical purchases using neural language models [5]. We represent each product p as a vector v_p in the user space; v_p(i) = k, k ≥ 0, if user i purchased product p k times during the history of a one-year time window. These product vectors are embedded into R\(^{100}\) using the FastText library\(^5\). Neighbor products in this domain are those purchased by ‘similar’ users.

An example of purchase similar products are the books “Becoming” by Michel Obama and “The Mother of Black Hollywood: a Memoir” by Jenifer Lewis (the autobiographies of two famous African-American women). Another example is “Pampers diapers size 0” and “Pampers diapers size 1” (note that the latter pair is also semantically similar).

Indirect relation based on purchase similarity (PurRel) between q and p is defined by the maximum cosine similarity between the embedding vector of p and the embedding vectors of all relevant products to query q:

PurRel(q, p) = \max_{p' \in R(q)} \cos(Embed(v_p'), Embed(v_p)) .

Figure 5 shows the indirect relation histograms for all irrelevant pairs (q, p), split into bins of equal size according to the semantic similarity score between p and q. Similar results were obtained by analyzing the indirect relation between products and queries in the context of purchase similarity, and are shown in Figure 6. According to the two measurements, the less related the product to the query, the lower the purchase/engagement level with it.

4.6 Data Analysis - Summary
To summarize, Table 4 covers the reasons discussed in this section that might explain user engagement with seemingly irrelevant search results.

5 SHOULD WE OFFER SEEMINGLY IRRELEVANT PRODUCTS TO CUSTOMERS?
Based on the results presented in Section 4, it seems that the rational decision for a search engine is to offer an objectively relevant product given the customer query, as purchase and engagement levels are expected to be higher for a relevant result than for an irrelevant result. While this is true in general, there are circumstances when this does not hold. In the following experiment we estimate the number of cases where offering an irrelevant product to the customer might be superior, in terms of purchase or engagement, to offering a relevant one. We sampled a set of 1.5K queries from our query log, where each query was associated with two different product items – one judged as relevant and one as irrelevant. For a given query q, and for each of its associated products p_i (i = 1, 2), we measured the average purchase and engagement levels over all occurrences of the (q, p_i) pairs in the log.

We measured the precision and engagement levels achieved over this set of queries, where each query in the set is associated with one of the two products according to a fixed selection policy. We experimented with five different policies: (1) Optimal selects the product
with the higher purchase or engagement level for the query, no matter its relevance; (2) Relevant always selects the relevant product; (3) Irrelevant always selects the irrelevant item; (4) Random selects the product randomly; and (5) Worst select the worst product in term of purchase/engagement level. Table 5 presents the normalized purchase and engagement levels achieved for each policy over all queries, normalized with respect to the Optimal policy.

<table>
<thead>
<tr>
<th>Policy</th>
<th>NPL</th>
<th>NEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Relevant</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>0.49</td>
<td>0.61</td>
</tr>
<tr>
<td>Random</td>
<td>0.58</td>
<td>0.64</td>
</tr>
<tr>
<td>Worst</td>
<td>0.17</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table 5: The normalized purchase and engagement levels when products are selected by different selection policies.

Obviously, the Optimal policy maximizes purchase and engagement levels, and Relevant is superior to Irrelevant, as expected. Random has a very high EL, not far from Relevant. Moreover, Relevant is far behind Optimal emphasizing that offering the relevant item is not always the right choice.

Looking further into this query set we found that for 26% of the queries offering the irrelevant product yields a higher purchase level than offering the relevant one. For 37% of the queries, offering the irrelevant product yields better engagement. These high ratios of queries, which lead to superior purchase/engagement level when associated with seemingly irrelevant items, demonstrate the need of search engines to consider various signals in addition to relevance, when deciding which product to offer to their customers.

### 5.1 The Relevance-Purchase Tradeoff

The significant number of cases when offering a seemingly irrelevant item has a higher purchase level than offering a relevant one raises the question whether the preference of an irrelevant product follows from the noisy nature of customer actions, or that we can optimize a ranking model for both objectives, namely, for relevance as well as for customer purchase or engagement level. For answering this question we experimented with a naïve approach that integrates two ranking models.

The ranking models were trained by a pairwise learning-to-rank process, using the LambdaMART algorithm [22]. The set of 1.5K queries with their associated products (one labeled relevant and one irrelevant) was used for training. Each \((q,p)\) pair was represented by a feature vector capturing similarity, relatedness, and behavioral signals. We used 10% of the data for hyper-parameter tuning (i.e., tuning the LambdaMART parameters – number of trees and leaves), and 10% for testing. The first model, REL, was optimized based on relevance labels only while ignoring any historical purchase or engagement signals. The second model, PUR, was trained on purchase level signals only ignoring the relevance labels. Similarly, ENG was trained on engagement level signals only. The model MIX\(_p\) is a fusion model that ranks the search results by linearly combining the normalized scores of REL and PUR\(^6\). Similarly, MIX\(_e\)

\[\text{MIX}_p(q,p) = \alpha \cdot \text{REL}(q,p) + (1 - \alpha) \cdot \text{PUR}(q,p)\]

\[\text{MIX}_e(q,p) = \alpha \cdot \text{REL}(q,p) + (1 - \alpha) \cdot \text{ENG}(q,p)\]

Figure 7 depicts the performance of the MIX models for different \(\alpha\) values in the range \([0, \ldots, 1]\). The x-axis represents purchase/engagement@1, i.e. the ratio of queries in the test set for which the top scored result was purchased/engaged. The y-axis represents relevance@1, i.e. the ratio of queries for which the top-scored result was relevant. The graph shows a clear tradeoff between relevance and purchase/engagement level. Optimizing for one of the objectives only, hurts the second objective and vice versa. Thus, we cannot ignore any of the signals in case we would like to optimize for both objectives. Moreover, a ranking model that optimizes for both objectives should be tuned according to the desired tradeoff between the objectives. In the following section we further elaborate on multi-objective optimization.

### 6 IMPROVING THE SEARCH EXPERIENCE

The findings presented in this work can be used in several directions for improving the customers experience in the product search domain. In the following, we describe two directions; one faces the customers directly while the other modifies the search engine ranking model.

#### 6.1 Search result justification

The first direction of improving the search experience targets the presentation of search results to customers. Customers’ experience might benefit from highlighting the positive properties of the results they are offered. This is particularly important when offering seemingly irrelevant results, and can be largely based on the hypotheses we presented. Mentioning that the product offered is related to a previous purchase, or that its price is lower than average, might enlighten the customer and contribute to a better experience. For example, consider the (artificial) case when a customer asks for

---

\(^6\)In order to integrate the model scores we normalized them to the range \([0,1]\) using max-min normalization approach.
‘Avocado’ and is offered ‘Banana’ instead. Positive explanations might look like:

- **Personalized preferences**: You usually buy Bananas while searching for Avocado.
- **Popularity**: Bananas are extremely popular now, hurry before running out!
- **Low price**: Bananas are on sale – only few left in stock!
- **Indirect Relation**: In general, people who look for Avocado also buy Bananas
- **Exploration mode**: Would you be interested to try our Bananas?

The choice of the explanation to be provided depends on the query \( q \) and the offered product \( p \). When \( p \) is clearly relevant to \( q \), a justification is not necessarily needed. In contrast, an irrelevant product \( p \) that previously appeared in the customer’s order or browse history, or that it is very popular, can be justified in a straightforward manner. This is also true for unspecified or exploratory queries, where we can safely assume that the customers will tolerate other types of offers. If \( p \) is related to \( q \), we should reveal the type of relation to justify our offer. Learning complex types of product relationship and their effect on the purchase decision is an open challenge that we leave for future work.

### 6.2 Multi-objective Optimization

To preserve customer trust and long-term engagement with the search system, reducing irrelevant results is an important goal from the search service perspective [11]. A multi-objective ranking model [15, 18, 21], which manages the apparent trade-off between products relevance and customers engagement, is expected to reduce the number of irrelevant offers, while preserving the engagement level.

The second direction we consider is based on multi-objective optimization and enhancement of the search engine’s ranking model with the behavioral signals discussed in this work. A classical ranking model is typically trained using a learning-to-rank (LTR) approach, in which a training example is a tuple \((q, p, l)\) consisting of a query \( q \), the offered product \( p \), and a label \( l \) representing a specific optimization objective. The goal of the trained model is to agree with the training examples on their associated labels [12]. In a multi-objective optimization setup, each such tuple contains several labels, e.g., the relevance label, the purchase label, and the engagement label of \( p \) to \( q \). The model is trained to be optimized with respect to all labels in hand.

In section 5, we demonstrated the output of a fusion model that linearly combines the output of two different ranking models: one optimized for relevance, and the other optimized for purchase/engagement level. Figure 7 showed a clear tradeoff between the two objectives, suggesting that a multi-objective optimization model is a promising direction for managing this tradeoff. Several approaches have been suggested for multi-objective learning to rank [9, 15, 18, 21]. A popular one is based on the lexicographic approach which sets preferences among the objectives to be optimized. It first optimizes for the primary objective, while considering the secondary objective in cases of ties [18]. Another popular approach is a linear combination of the objectives. The ranking model is trained based on an aggregated label which is a linear combination of all labels, weighted according to their relative significance.

Moreover, taking into account different factors and the derived signals that have been mentioned in this work, could allow us to better tune our ranking models and to trade between relevance and purchase/engagement levels. Examples of such signals could be to take into consideration the query intent (buy vs exploration), or the customer’s tolerance level (approximated by the query length) when optimizing the dual relevance and purchase/engagement objective. Exploring multi-objective optimization solutions while using these signals is a future direction of our work.

### 7 SUMMARY

In this work, we analyzed the interesting phenomenon in voice product search, in which customers purchase or engage with seemingly irrelevant search results. We demonstrated that this is a significant type of user behavior that requires attention, and addressed several hypotheses for the reasons behind it.

Looking at the source of the offered products, we observed that seemingly irrelevant offers based on past-purchases or on popularity have high purchase/engagement levels. Looking at the query intent, we observed that the level of purchase and engagement with irrelevant results is higher in exploration mode, as the customer is not necessarily in the context of an explicit purchase need. We continued by demonstrating that purchase and engagement levels depend on the query specificity, are biased towards low-price offers, and correlate with the indirect relationship of the product with the query.

We then analyzed the circumstances in which it might be reasonable to offer seemingly irrelevant products to our customers. We demonstrated that it would not be optimal for a voice search engine to consider only product relevance for result selection, as in a significant number of cases, offering an irrelevant product yields a higher purchase level. We demonstrated that there exists a tradeoff between relevance and purchase levels, emphasizing the need to consider both objectives for optimizing search performance.

We concluded with a discussion on two possible directions for future research, towards improving user’s experience. The first direction considers optimizing the ranking model, managing the apparent trade-off between products relevance and the customers engagement level. The second direction targets the presentation of the search results to customers, which might benefit from highlighting the positive properties of the results they are offered. This is particularly important when offering irrelevant results, and can be largely based on the hypotheses we presented.

Many future research directions are yet to be explored. In this work we only considered the implicit purchase and engagement signals as indicators for user satisfaction. It would be interesting to explore explicit user feedback on seemingly irrelevant results, as well as how the tolerance with irrelevant results vary across users and affects their purchase behavior. Moreover, we only analyzed one-search sessions in this work. It would be interesting to extend this study to multi-search sessions in product search and the temporal aspects of user engagement with seemingly irrelevant results.
Why Do People Buy Seemingly Irrelevant Items in Voice Product Search?

REFERENCES


