JDivPS: A Diversified Product Search Dataset

Zhirui Deng
Zhicheng Dou
Yutao Zhu
zrdeng@ruc.edu.cn
dou@ruc.edu.cn
yutaozhu94@gmail.com
Gaoling School of Artificial Intelligence,
Renmin University of China
Beijing, China

Xubo Qin
Pengchao Cheng
Jiangxu Wu
Hao Wang
qratosone@live.com
chengpengchao3@jd.com
wujiangxu@jd.com
wanghao66@jd.com
JD.com, Inc.
Beijing, China

ABSTRACT

The diversification of product search aims to offer diverse products to satisfy different user intents. Existing diversified product search approaches mainly relied on datasets sourced from online platforms. However, these datasets often present challenges due to their restricted public access and the absence of manually labeled user intents. Such limitations may lead to irreproducible experimental results and unreliable conclusions, restricting the development of this field. To address these problems, this paper introduces a novel dataset `JDivPS` for diversified product search. To the best of our knowledge, `JDivPS` is the first publicly accessible dataset with human-annotated user intents. The dataset is collected from JD.com, a major Chinese e-commerce platform. It includes 10,000 queries, around 1,680,000 unique products, and an average of 10 human-labeled user intents for each query. We have extensively evaluated several diversified ranking models using the `JDivPS` dataset. The results of these models are recorded and presented, serving as a valuable benchmark for future research. More details about the dataset can be found in https://github.com/DengZhirui/JDivPS.

CCS CONCEPTS

• **Information systems** → **Information retrieval diversity**.

KEYWORDS

Product Search, Diversification, Dataset

ACM Reference Format:
Zhirui Deng, Zhicheng Dou, Yutao Zhu, Xubo Qin, Pengchao Cheng, Jiangxu Wu, and Hao Wang. 2024. JDivPS: A Diversified Product Search Dataset. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626772.3657888

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657888

1 INTRODUCTION

In web search, it is common for users to submit short queries, which are often ambiguous or vague [10, 14, 24, 25]. Similar phenomenons have been observed in product search [6, 21]. Customers often represent their shopping intents in broad terms without detailing specific requirements [19]. For example, a search for "outdoor barbecue" may not specify whether the customer is looking for barbecue grills, accessories, or related outdoor furniture. To effectively satisfy users' different intents, e-commercial search engines need to return a diverse set of products. This approach can not only enhance users' satisfaction by aligning more closely with their vague or multifaceted needs but also play a crucial role in distributing consumer traffic across a wider range of products. A search result set that is too narrow or homogeneous can adversely affect both user engagement and commercial profits. To address these challenges, diversified product search strategies have been recognized as a pivotal solution.

Given a query and an initially ranked list of products based on relevance scores, diversification models aim to adjust the ranking to balance relevance and diversity. The inclusion of a variety of products in search results is not merely a strategy to satisfy the different needs of a broad user base but also to address a user's inherent interest in discovering diverse options. It is important to note that, within the structure of commercial product search engines, the diversification model we discuss in this paper is placed between the retrieval and the re-ranking stage. This placement is deliberate to ensure that diversification is considered before finalizing the products presented to the user. Moreover, this study focuses on a non-personalized approach to product search diversification, as personalized product search [17, 18] requires additional considerations, which are beyond the scope of this paper.

For the diversification of both web search [9, 16, 22] and product search [6, 21], the research community is struggling with the lack of datasets equipped with diversity intent annotations. The widely used TREC WebTrack 2009-2012 dataset for web search includes only 198 queries with diversity intent annotations [16].[1] It is even more severe for e-commerce search since **there are no publicly available datasets including intent annotations**. To navigate this constraint, previous researchers have had to use the sub-category datasets from Amazon, initially designed for general

[1] https://trec.nist.gov/data/web09.html

recommendation purposes. The absence of human-labeled intents for each query in these datasets means that researchers are compelled to infer consumer intents based on product categories.

We argue that using product categories as proxies for consumer intents does not accurately reflect the complexity of real-world scenarios on e-commerce platforms. First, e-commerce platforms design their product category systems primarily to facilitate product management for sellers, not to reflect consumer recognition accurately. Consequently, these taxonomies may lead to terminal-level categories that lack the granularity needed to precisely represent user intents. For example, a terminal-level category such as "Computer & Tablet" on Amazon.com insufficiently differentiates between PCs and tablets, which for most consumers, serve distinct needs and are not interchangeable. Second, the intent behind the queries is multifaceted, including categories, brands, or other attributes of the products. However, due to the absence of real queries in the Amazon dataset, researchers are forced to rely on existing categories or attributes as pseudo queries. This approach simplifies user intents into a single dimension, failing to capture the multifaceted nature of consumer search behavior. Finally, diversification in product search must strike a balance between enhancing the user experience and optimizing commercial profit. It is impractical to prioritize long-tail intents with limited commercial value on par with those having substantial commercial potential intents.

In this paper, we release a novel dataset, JDivPS (JD Diversified Product Search dataset), derived from the JD.com platform. JDivPS has the following characteristics.

• **Data Scale**. The dataset contains about 10,000 real queries and about 1,680,000 unique products, including a testing set of 1,000 queries with human annotations.

• **Multiple Intent Types**. We classify queries into three intent types: fine-grained product categories (referred to as *Fine-grained Categories*), brands, and attributes. Distinctly, rather than relying solely on the existing product categories in product metadata, our dataset utilizes actual product related phrases submitted by users (*e.g.*, "gaming laptop") to more accurately capture fine-grained category intents.

• **Initial Ranking List**. Each query is associated with an initial ranking list of product items. The relevance between intents and items is determined by a BERT-based relevance matching model currently in use on the JD.com platform [15].

• **Intent Annotations**. We use the relevance model in the platform [15] to annotate the intent coverage of queries and products. For the testing set, both the query-intent and intent-product relevance labels have been rigorously judged by human annotators.

• **Product Metadata and Query Suggestions**. In addition to queries, intents, and candidate product lists that are essential to product search diversification, we provide supplemental product metadata and query suggestion data. These data can help researchers develop advanced product ranking and intent mining algorithms.

• **Pretrained Models**. Besides the dataset, we also provide a BERT model pretrained with about 10 million product titles, and another BERT model distilled from a high-performance teacher model [15] to compute the relevance between queries and product titles. Due to the lack of training data, only a few existing diversification approaches [23] tried to leverage fine-grained token-level matching signals. Our dataset and models can significantly assist researchers in implementing diversification approaches leveraging token-level interactions. We will later release more pretrained and fine-tuned model checkpoints to support the research based on JDivPS dataset.

JDivPS is licensed under CC BY-NC-SA 4.0.[2] Key information in the dataset is anonymized to protect user privacy. We already make the dataset with detailed documentation publicly available via https://github.com/DengZhirui/JDivPS including some tools to help process the dataset.

With the data, we implement a variety of existing diversification approaches based on JDivPS and evaluate their effectiveness. Our experimental findings suggest designing specific diversification strategies for product search, highlighting the limitations of directly applying web search diversification methods.

Overall, JDivPS has the following advantages to benefit the research community of product search diversification:

(1) To our knowledge, JDivPS is the first dataset specialized for product search diversification. With its substantial query volume that is approximately twenty times that of TREC WebTrack, JDivPS can reduce the risk of overfitting and promote the development of more sophisticated diversification algorithms.

(2) JDivPS is derived from real user interactions on JD.com, a leading Chinese online shopping platform. Real queries, rich product metadata, and high-quality query suggestions make it possible to develop industry-level diversified product search methods.

(3) We provide human-labeled intents for each query and detailed intent-item annotations for the testing set. By employing phrases that reflect fine-grained categories, JDivPS ensures a precise representation of actual product intents, with all phrases verified by human annotators to guarantee accuracy in reflecting user intents.

2 RELATED WORK

Datasets adopted by existing diversified product search works [6, 11, 12, 21] can be categorized into data crawled from commercial websites and the Amazon dataset.[3]

Data collected from online shopping platforms can reflect real user behavior, but existing datasets are non-public, and researchers [6, 21] did not illustrate the information and the distribution of these datasets. The non-disclosure of data prevents the dataset from being reused and forces the subsequent works to reconstruct a new dataset which restricts the advancement of this field. Moreover, these datasets utilized the re-ranking results provided by existing shopping platforms which was affected by the re-ranking model. Although some products are needed by users, they cannot be ranked at the top by the re-ranking model due to short user queries or problems with the fine-ranking model, and cannot appear in the initial ranking. This will cause the data to be inherently biased. Furthermore, such datasets adopted product categories as a proxy of real user intents. However, the real user intent includes not only product categories but also the product's brand and other attributes.

As for the Amazon dataset [20], it is a recommendation dataset and only contains user ratings and reviews on products and the

[2]CC BY-NC-SA 4.0 license, https://creativecommons.org/licenses/by-nc-sa/4.0/
[3]Amazon dataset: http://jmcauley.ucsd.edu/data/amazon/links.html

metadata of the products. Therefore, previous product search methods [1–3, 18] manually construct pseudo queries by concatenating the categories of the product. However, in actual search scenarios, users are involved in more than just categories. Therefore, the constructed fake queries cannot reflect the true distribution of user behaviors. Besides, diversified product search requires human-annotated user intents under each query and the correspondence between products and intents to evaluate the performance of the model. However, the Amazon dataset does not have these annotated data and is not suitable for the diversified product search. Therefore, a new dataset is urgently needed for the diversified product search.

Different from previous datasets, in this work, we construct a diversified product dataset based on the real user search log on JD.com which is a large online shopping platform. We also manually annotated user intents and the intent coverage of products. To the best of our knowledge, we are the first diversified product search dataset with human-annotated real user intents.

3 THE JDIVPS DATASET

Diversified product search aims to satisfy the different needs of users by presenting a broad spectrum of product options. Datasets adopted by previous approaches are not publicly accessible datasets without manual user intent annotations. Although some studies have attempted to utilize the Amazon dataset to explore product search diversification [13], this dataset does not contain actual user queries and assumes product categories as proxies for real user intents which actually cannot adequately represent users' real information needs. Given these constraints, current datasets cannot accurately reflect multifaceted real user intents which often include product types, brands, and other specific attributes. This limitation potentially renders the developed, tested, and evaluated methods less effective or irrelevant in practical applications. Therefore, it is essential to develop and release a new dataset that is rooted in real user behaviors and includes manually annotated user intents.

3.1 Definition of Query Intents

For large e-commerce platforms, it is common for users to initiate searches that extend beyond products, such as looking for specific stores or services, like "JD Supermarket" or "secondhand products recycling". To align with the application scenarios of large e-commerce platforms, we refine the concept of "user intents" to specifically denote users' needs for particular types of products, excluding non-product related queries. Within our dataset, we categorize product-related query intents into three dimensions:

(1) **Fine-Grained Category** denotes a precise word or phrase that identifies an atomic type of product, such as "tablet" or "gaming laptop". On e-commerce platforms, a terminal-level product category may cover a variety of such fine-grained categories, *e.g.*, both "tablet" and "gaming laptop" belong to the terminal-level category of "Computers & Tablets".

(2) **Brand** indicates the brand name of a product, *e.g.*, "Apple" or "Lenovo". Brands serve as a form of human-labeled metadata that sellers provide to describe their products.

(3) **Attribute** describes specific properties of products, including color, model, size, etc. These details are typically furnished by sellers in the product metadata to help define the product's specifications.

Based on these definitions, we divide the queries into **Category Ambiguous Queries** and **Category Clarified Queries**. Category ambiguous queries, such as "apple" for a specific brand or "outdoor barbeque" for an application scene, generally indicate a need for identifying a **fine-grained category**. Conversely, category clarified queries contain phrases that specify a need for a particular product type, directing the associated intents towards **Brand** and/or **Attribute** dimensions (*e.g.*, gaming laptop). It is important to note, however, that even queries that seem ambiguous may still carry underlying intents related to brand or attribute specifications.

3.2 Fundamental Data

Based on the definition of user intents mentioned above, we develop a new dataset `JDivPS` to tackle the problems in existing datasets and facilitate research in the field of diversified product search. Our dataset `JDivPS` is derived from real user interactions on JD.com, a leading Chinese online shopping platform. The dataset construction process is illustrated as follows.

3.2.1 User Query Collection. We collect all user queries on the online shopping platform on a specific date, November 30, 2023. Recognizing that raw query data may include low-quality or redundant entries, we implemented a rigorous filtering process to enhance the dataset's quality. Initially, we exclude queries not in Chinese, such as numerals or abbreviations irrelevant to product searches (*e.g.*, "13" and "jd"). We further remove duplicate entries and queries comprising solely system-related terms like "live broadcast". To focus on queries with significant user engagement, we retain those appearing more than 20 times within one month, thus excluding less frequent (long-tail) queries. Additionally, we applied a category prediction model as described in [29] to each query, discarding those associated with over 50 category predictions, which typically indicate non-product queries (*e.g.*, "flagship store"). Furthermore, utilizing the platform's existing retrieval pipelines, we identify relevant products for each query. Each product's relevance to the query is assessed using a deep model [15]. Queries linked to fewer than 20 products scoring a relevance above 0.5 are filtered out to remove low-quality queries or those not necessitating diversification. After this preprocessing phase, we randomly selected approximately 10,000 queries to construct our dataset.

3.2.2 User Intents Collection. Considering that a user's click or purchase behavior under a specific query may indicate their real intents, this study places a greater emphasis on purchase behavior to derive real user intents. This decision is grounded in the observation that clicks, while indicative, can sometimes result from accidental actions and may introduce noise. At first, we collect users' purchasing activities on JD.com over the period from November 1, 2023, to November 30, 2023. These data are organized in chronological order using timestamps to track each user's historical interactions. Any users whose queries do not match those in our dataset are excluded.

For each query in our dataset, we aggregate all products purchased in response to that query throughout the observed period. Recognizing that users often refine their searches with more precise queries following an unsuccessful broad initial search, we also identified and incorporated these refined queries from the users'

purchase histories. A sliding window approach is employed, where a window of ten subsequent interactions is considered for each query in our dataset. If a subsequent query within this window shares common terms with the original query, the products purchased under these refined queries are also considered as the user-desired products under the original query.

This comprehensive aggregation aims to cover the search intent of all users under each query. However, acknowledging the potential for noise within the purchase data, a series of preprocessing steps are conducted to refine the collection of products associated with each query, thereby obtaining the initial user intents. Concretely, products whose names do not overlap with the terms in the query are first filtered out. Based on the assumption that search intents on e-commerce platforms are typically directed towards specific products, product characteristics such as brands, models, sizes, colors, and fine-grained categories of these products are used as proxies for initial user intents. Then, redundant intents, those reduced to mere numerical form, or directly echoing the query are eliminated. For broad queries like "cellphone", user intents are further filtered to offer more targeted product options. For highly specific queries (*e.g.*, retaining a singular focus on brand or product), all corresponding intents are preserved; otherwise, emphasis is placed solely on retaining brand and product terms as the intents.

3.2.3 Product Metadata Collection. In our dataset, we systematically categorize product information into several key metadata fields: name, brand, category, and attributes (including fine-grained characteristics like the model, size, and color). For instance, for a product labeled "Huawei cellphone X5", "X5" is identified as the attribute. In addition to these descriptive metadata, we incorporate statistical features for each product, such as Unique Visitor (UV) count, Page View (PV) number, and Click-Through Rate (CTR). These metrics are calculated independently of any specific queries or user profiles, serving as indicators of a product's inherent popularity and commercial appeal. A product with a higher UV or PV indicates greater visibility on the platform, whereas a higher CTR suggests a stronger ability to engage users and encourage them to click through for more information.

3.2.4 Data Partition and Intent Judgement. JDivPS contains 10,000 queries. We randomly sample 10% of the queries (1,000) in our dataset as the test set, and the left 9,000 are used for training purposes. Each query is associated with an initial ranking of 200 products. We randomly sample 10% of the queries in our dataset as the test set.

We leverage the platform's established retrieval pipelines to judge if an intent is relevant to a query. Given a query and an intent, **we formulate an "extended query" by concatenating the original query with the intent**. This enhanced query is used in a two-phase retrieval process. Initially, products are identified using both sparse and dense retrieval techniques, followed by an evaluation of each product's relevance through a BERT-based matching model. It is important to note that products that are either unpopular or not currently available (out-of-stock) are excluded before the relevance assessment phase. Further information on the relevance model is available in [15].

Given an extended query and its corresponding products, the relevance model will judge if the products are relevant to all the components in the extended query. As a result, a product must satisfy the phrases in both query and intent to achieve a higher relevance score. **If an intent is relevant to a query, the products retrieved by the "extended query" must be highly relevant.** Here "highly relevant" means the products have got their relevance scores higher than 0.9. If the retrieval pipelines fail to yield any products meeting this high relevance criterion, it suggests the intent does not align with the query, leading to the exclusion of that intent from further consideration. All the query-intent relations in both train and test sets are annotated initially by the relevance model mentioned above. For the test set, the query-intent and intent-product relations are double-checked by human annotators. More details about human annotation will be described in Section 3.3.

Table 1: Attribute and description of the data.

Subset	Attribute	Description
initial ranking	query	query's anonymized term ids
	product_id	anonymized id of a product in the initial product list
	relevance_score	relevance of the product to the query
	tf_idf_name	tf-idf score of the product's name
	tf_idf_category	tf-idf score of the product's category
	tf_idf_brand	tf-idf score of the product's brand
	bm25_name	BM25 score of the product's name
	bm25_category	BM25 score of the product's category
	bm25_brand	BM25 score of the product's brand
	UV, PV, CTR	UV, PV, and CTR score of the product
product metadata	product_id	the product's anonymized id
	product_name	the product's anonymized term ids
	brand_name	the product brand's anonymized term ids
	category_name	the product category's anonymized term ids
	attribute	the product attribute's anonymized term ids
	size	the product size's anonymized term ids
	color	the product color's anonymized term ids
query suggest.	query	query's anonymized term ids
	query_suggestion	anonymized term ids of the suggestion
intent-product relevance	query	query's anonymized term ids
	intent	anonymized term ids of a user intent
	product_id	anonymized id of a product in the initial product list
	relation (0/1)	relevance of a product to the intent

3.2.5 Initial Ranking Construction. From the pool of products with scores of highly relevant, we select a random subset of p products as positive examples. In alignment with the objectives of diversified product search, we incorporate some products that are less relevant to the query to the initial product list. This is achieved by randomly choosing q products from the retrieved set and including them in the preliminary product list. After obtaining the initial product list, the ranking is refined using the original query without any specific intent through the aforementioned query-product relevance model.

3.2.6 Query Suggestion Collection. To approximate the actual user intents, which are not directly available for online ranking, our dataset includes query suggestions derived from JD.com (called *subtopics*), serving as proxies for these intents. For each query, we

collect suggestions under the query on JD.com between November 1 and November 30, 2023. From this collection, we randomly selected 10 suggestions per query to act as subtopics. These subtopics are intended for use by diversification algorithms as proxies for the real user intents associated with each query.

3.2.7 Anonymization. Due to the commercial confidentiality rules of the platform, we are not allowed to provide the original data. So we have anonymized JDivPS to uphold confidentiality requirements. Textual data, including queries, suggestions, product titles, brands, categories, attributes, sizes, and colors, are tokenized into integer IDs using a private tokenizer. To support research with this anonymized textual data, we provide a BERT model pre-trained on the same tokenization scheme and a BERT-based relevance matching model distilled from the platform's primary model [15]. Statistical features such as UV, PV, and CTR have been normalized, with UV and PV values mapped to a 0-10,000 scale, preserving the integrity of product popularity and commercial value comparisons.

The attributes and descriptions of our JDivPS dataset are illustrated in Table 1. Comprehensive information about JDivPS, including access instructions, is available in our repository.

3.3 Intent and Product Relevance Annotation

As we mentioned in Section 3.2.4, for both train and test sets the query-intent and intent-document relations are judged by the relevance model of the platform initially. For the queries in the test set, manual annotation is conducted to ensure the correctness of the judgment given automatically by the model. In this section, we describe the annotation guidelines, using examples of real user queries and product titles, translated from Chinese to English, to illustrate our annotation principles clearly.

3.3.1 Definition of Product Relevance. Following the platform's established relevance matching guidelines, we categorize a product as "relevant" to a query if all the query's phrase components are reflected in the product's name or its metadata. "Matching" is defined as instances where the query phrases either exactly or semantically correspond to phrases associated with the product. Moreover, when a query specifies a fine-grained category, the product must precisely belong to that category to be considered relevant.

• **Example:** For the query "down jacket", a product named "Lightweight Down Jacket with Hood, Outdoor Fall and Winter Short Coat for Men" is deemed relevant. Conversely, a product named "Women's Down Jacket Sleeve Covers, White Arm Warmers, Cotton Office Work Sleeves, Dirt and Stain-Proof, Arm Guards with Cuffs" will be marked irrelevant because its primary category, "sleeve cover", diverges from that of the query "down jacket", despite the matching phrases.

3.3.2 Intent Annotations. Our intent annotations are based on the aforementioned relevance criteria. Annotators assess the correlation between queries and each associated intent. They use the concatenation of a query and an intent as an "extended query" to search the platform. Ignoring advertised products, which often have lower relevance, the presence of at least one relevant product in the search results qualifies the intent as relevant to the query.

• **Example:** For the query "Chinese Knot" with the intent "Festival Decorations", searching "Chinese Knot Festival Decorations" might return relevant products like "2024 Dragon Chinese Knot Ornament New Year Decoration for Entryway, Living Room Spring Festival Lucky Character Hanging Decor, Large Size". Finding such products confirms the intent's relevance to the query.

A negative case is a query "outdoor barbeque" with a mismatched intent "medical box". The top relevant product can be "Car Portable Insulated Cooler Box, Thick Aluminum Foil Ice Pack, Ice Bag for Freshness, Outdoor BBQ and Takeout Food Preservation Box.". The product is relevant to the query, but it's irrelevant to the intent which will be annotated as irrelevant to the query and excluded.

3.3.3 Intent-Product Relevance Annotations. After filtering out noisy intents, annotators evaluate the relevance between products and intents. For each intent, they judge the relevance between the products and the extended query mentioned above. A relevant product should be relevant to both the intent and its corresponding query.

• **Example:** Continuing with the query "Chinese Knot" and the intent "Festival Decorations", a product like "2024 Dragon Chinese Knot Ornament New Year Decoration" that matches both the query and the intent in its name, and aligns with the query's fine-grained category, is labeled relevant.

Note that during the above annotation, each sample is reviewed by three annotators, and a majority vote determines the final label.

Table 2: Statistics of the JDivPS dataset

Statistics	Value
#Total queries	10,000
#Training queries	9,000
#Test queries	1,000
#Products per query	200
#User intents per query in the training set	10.68
#User intents per query in the test set	14.83
#User intents per query in the entire set	11.26
#Query suggestions per query	10

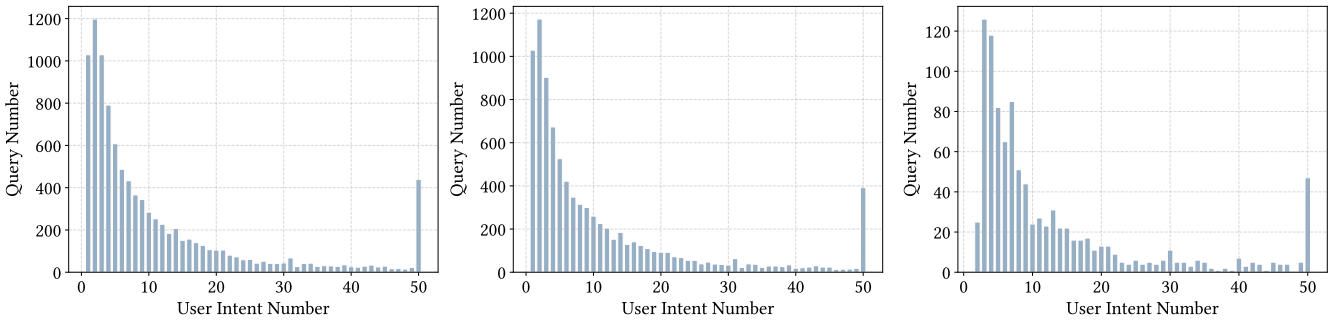
4 DATASET ANALYSIS

The overall statistics of JDivPS are shown in Table 2. Recall that the training set's intents are derived automatically from user logs. For the testing set, user intents are manually labeled, enhancing the reliability and depth of evaluation.

Next, we will investigate various aspects of JDivPS in depth to comprehensively show our dataset.

4.1 Query Analysis

The analysis of user queries in JDivPS reveals a diverse spectrum of intent granularities. For example, users searching for the brand "Huawei" may refine their search to "Huawei mobile phone" or "Huawei laptop", indicating the intent being fine-grained categories or specific brands. On the other hand, a precise query such as "iPhone 15 256g" suggests that the user's intent leans more towards particular product attributes, rather than the category or brand, which are already indicated in the query. This variance in query specificity is reflected by the number of intents per query, with



(a) Entire Set (b) Training Set (c) Testing Set

Figure 1: The distribution of user intents number of queries.

Table 3: The distribution of intent types in JDivPS.

Dataset	Fine-Grained Category	Brand	Attribute
Training	46.8%	52.4%	0.8%
Test	53.1%	46.5%	0.4%
Entire	47.7%	51.6%	0.8%

Table 4: Popularity Metrics of the Products

Item	Unique Visitor (UV)	Page View (PV)	CTR
Overall	244	226	0.034
Positive	254	233	0.037
Negative	227	213	0.029

broader queries typically associated with a higher number of intents compared to more clear queries. To quantitatively depict this relationship, we analyzed the distribution of intent numbers across queries in JDivPS, as illustrated in Figure 1. To maintain visual clarity and consistency in our analysis, we have set a cap on the displayed intent count at 50 for queries exceeding this threshold.

The analysis highlights a similar distribution of intent quantities across both training and testing datasets, showcasing a decrease in query frequency correlating with an increase in the number of user intents. This pattern indicates that although our approach does not explicitly incorporate a popularity metric or "heat factor" to extract user intents, it implicitly accounts for the relative prominence of these intents by analyzing user logs. Consequently, infrequently searched intents are naturally sifted out as long-tail intents.

4.2 Intent Analysis

There are three types of user intents in our datasets, *i.e.*, fine-grained category, brand, and attribute. As shown in Table 3, analysis of these user intents reveals significant insights into search behavior:

(1) **Predominance of Fine-Grained Categories and Brands:** The analysis indicates a strong user preference for searching within broad categories or brands, with only a small fraction (0.8%) of user intents relating to product attributes in the training set, and an even lesser proportion in the testing set. This pattern highlights a general tendency among users to initiate searches with a broader scope, rather than seeking out products based on specific features.

(2) **Brand Preference:** In both the training and entire sets, brand-related intents constitute the majority, accounting for 52.4% and 51.6% of intents. This trend aligns with typical consumer behavior, where users are often uncertain about the desired brand after determining the product type they are interested in.

4.3 Product Popularity Analysis

We also analyze the popularity distribution of products in our dataset. Table 4 shows the average UV, PV, and CTR values of all products in the dataset. In the table, "Overall" denotes the average metrics across all products, "Positive" denotes the metrics of products with positive intent annotations, and "Negative" denotes the metrics of the rest irrelevant products. UV and PV are floored or ceiled to integers. We can observe that products with positive intent annotations generally exhibit higher UV, PV, and CTR. This observation supports the notion that our sampling strategy (introduced in Section 4.1) effectively identifies popular, rather than long-tail intents. It highlights the importance of balancing relevance and diversity in search results, alongside optimizing for user experience and commercial viability. The popularity features provided in our dataset offer a realistic snapshot of consumer demand on e-commerce platforms. Leveraging these metrics can enhance the performance of diverse ranking algorithms, ensuring they align with both user needs and commercial objectives.

5 BENCHMARK EXPERIMENTS

5.1 Benchmark Approaches

In this section, we describe our benchmark approaches in detail.

5.1.1 Non-diversified Baselines. Our benchmark includes the following baselines with no diversification:

REL is the initial ranking mentioned in Section 3.2.4, where all products are ranked only with the relevance scores.

UVCTR is another ranking approach based on both relevance and product popularity. Inheriting the spirit of the actual ranking strategy in the platform, we propose a simplified ranking function denoted as follows:

$$S_{q,p} = (\text{Rel}(s_{q,p}) * 10000 + \text{UV}_p \times \text{CTR}_p) \div 50000.$$

Here $S_{q,p}$ is the ranking score of product p with query q, $s_{q,p}$ is the relevance score of $< q, p >$, UV_p and CTR_p are the UV and CTR of p. Rel is a mapping function that converts the relevance score into a 5-level integer label from 0 to 4. The UVCTR ranking function will map all the products into 5 relevance levels, in each relevance level the ranking positions of products are determined mainly by $\text{UV}_p \times \text{CTR}_p$. The relevance label is timed by 10000 since the maximum value of UV after normalization is 10000, ensuring that a relevant product must get higher scores than an irrelevant one. This ranking strategy can promote popular products without compromising their relevance. Notice that the approach of UVCTR is a re-ranking strategy, **it is not another initial ranking baseline**.

5.1.2 Diversification Baselines. There are limited works for diversification in product search and previous works [4, 6] mainly adopt MMR framework to design models. Besides, diversified product search is similar to search result diversification in terms of task definition and dataset structure, except that it re-ranks products instead of documents. Therefore, we also utilize some latest search result diversification models [22, 27] as benchmark models. We omit other supervised approaches with greedy selection [16, 26] due to time limitations.

MMR [4] iteratively selects the next product combining the relevance between the query and the current product and the divergence between the current product and the selected products. Since MMR is a linear combination of initial ranking scores and novelty scores, we provide different kinds of MMR implementation based on different initial ranking baselines: $\textbf{MMR}_{\text{REL}}$ and $\textbf{MMR}_{\text{UVCTR}}$ with REL and UVCTR as the relevance model, respectively. We also provide implementations based on different phrases in the product metadata. More details can be found in Section 5.5.

DESA [22] leverages the attention mechanism to simultaneously evaluate the documents' diversity.

DALETOR [27] derive differentiable diversification-aware losses to approach the optimal ranking.

5.2 Evaluation Metrics

To rigorous evaluate the performance of baseline models, similar to previous diversified product search models [6], we also leverage evaluation metrics first adopted by search result diversification models [9, 22, 26], including α-nDCG@{10, 20} [7], ERR-IA@{10, 20} [5], NRBP [8], P-IA@{10, 20}, and S-rec@{10, 20} (*i.e.*, Subtopic Recall [28] to evaluate the performance of models on our datasets.

5.3 Implementation

For the supervised approaches, we train the model on our training set and calculate the evaluation metrics on our test set. We use a BERT encoder pre-trained for text representations [4] to generate the embeddings for all the products. We concatenate all the textual metadata of each product to generate the product embeddings.

Following the instructions of [22], we implemented an implicit variation of DESA without subtopics with a 3-layer encoder only to simplify the problem. The hidden state of the encoder is 256 and the head number for multi-head attention is 16. We use the list-pariwise sampling approach [16] to generate training data pairs for DESA. The max length of the context sequence is 50, and only the positive-negative pairs with an α−nDCG metric distance larger than 0.1 are preserved. For DALETOR, we set the batch size, epoch, and learn rate as 8, 1, and 0.01, respectively.

[4]https://huggingface.co/WangZeJun/simbert-base-chinese

5.4 Result and Discussion

Experimental results are shown in Table 5. Surprisingly, we find that MMR with UVCTR as the relevance model (*i.e.*, $\text{MMR}_{\text{UVCTR}}$) is the best approach with the top performance, outperforming most of those non-diversified and diversified baselines. Besides, the performance of both DESA and DALETOR are inferior to $\text{MMR}_{\text{UVCTR}}$. The performance of DESA is extremely poor. A possible reason is that the list-pairwise sampling strategy used in our experiments was originally designed for the small-scaled TREC Web Track dataset, it is ineffective and unstable for large datasets. A carefully designed sampling strategy for the large-scaled `JDivPS` dataset may be necessary. We leave this as a future work. Another possible reason may be the sentence-based representation encoder is unsuitable to generate a high-quality representation for product titles. Details will be discussed in Section 5.4.2.

5.4.1 UVCTR v.s. REL. In real-world e-commerce platforms, the relevance matching components play a different role compared with the components in web search. In product search, relevance matching happens in the early stage of the ranking pipelines. Based on the results of relevance matching, there will be further ranking stages (*e.g.*, CTR estimation with user profiles [30]). As a result, the relevance scores will not be discriminated enough to distinguish two products. For example, a product with a relevance score of 0.95 is almost identical to another product with a score of 0.93. As a result, ranking solely based on the relevance score may not lead to a good performance.

Compared with **REL**, **UVCTR** can achieve the balance between the relevance and popularity of the products. As we mentioned in Section 5.1.1, **UVCTR** leverages different levels' relevance score: at each relevance level, the popularity of the products determines the ranking scores most of all. Popularity can help distinguish products covering different user intents when there is little differentiation of the products' relevance scores, leading to better performance.

5.4.2 The Representation of Products. As we mentioned above, we leverage a pre-trained BERT encoder via some self-supervised approaches (*e.g.*contrastive learning) for generating high-quality sentence representations. Both MMR and other supervised approaches depend on the representations to model the novelty of different products. In a similar vein, in web search, web pages are divided into passages and produce representations through an encoder. Based on those representations, the representation-based diversification approaches [9, 16, 22, 26, 27] can measure the diversity of documents effectively.

However, **a product title is a combination of phrases instead of a sentence**. In a product title there may be multiple different types of phrases to describe a product, *e.g.*central product category, the synonym of category, features description, highlight, and application scenario. Take the product "Women's Down Jacket Sleeve Covers, White Arm Warmers, Cotton Office Work Sleeves, Dirt and Stain-Proof, Arm Guards with Cuffs" mentioned in Section 3.3 as an example, we list all the component phrases and their types in

Table 5: Overall performance of benchmark models. The best results are marked in bold.

Model	α-nDCG@10	α-nDCG@20	ERR-IA@10	ERR-IA@20	NRBP	P-IA@10	P-IA@20	S-rec@10	S-rec@20
REL	.0810	.1090	.0293	.0351	.0206	.0155	.0154	.1252	.2251
UVCTR	.2064	.2898	**.0838**	.1001	.0621	**.0456**	.0499	**.3099**	**.5318**
MMR_{REL}	.0874	.1164	.0312	.0374	.0226	.0164	.0164	.1283	.2302
MMR_{UVCTR}	**.2154**	**.2991**	.0836	**.1001**	**.0623**	.0454	**.0500**	.3077	.5312
DESA	.0399	.0364	.0224	.0224	.0233	.0035	.0017	.0350	.0350
DALETOR	.0797	.1311	.0254	.0341	.0175	.0124	.0170	.1202	.2864

Table 6: Results of Ablation Study.

Model	α-nDCG@10	α-nDCG@20	ERR-IA@10	ERR-IA@20	NRBP	P-IA@10	P-IA@20	S-rec@10	S-rec@20
MMR_{REL_name}	.0874	.1164	.0312	.0374	.0226	.0164	.0164	.1283	.2302
MMR_{REL_brand}	.0860	.1157	.0304	.0367	.0219	.0161	.0165	.1246	.2264
MMR_{REL_cate}	.0834	.1121	.0307	.0366	.0226	.0160	.0159	.1237	.2228
MMR_{UVCTR_name}	.2154	.2991	.0836	**.1001**	**.0623**	**.0454**	**.0500**	.3077	**.5312**
MMR_{UVCTR_brand}	.2157	**.2993**	**.0837**	**.1001**	**.0623**	**.0454**	**.0500**	**.3096**	**.5312**
MMR_{UVCTR_cate}	**.2159**	.2990	**.0837**	.1000	.0622	**.0454**	**.0500**	.3087	.5303
MMR_{EMB_name}	.0652	.0876	.0216	.0260	.0159	.0110	.0113	.0855	.1611
MMR_{EMB_brand}	.0686	.0977	.0201	.0258	.0141	.0106	.0133	.0887	.1821
MMR_{EMB_cate}	.0591	.0769	.0185	.0220	.0138	.0092	.0092	.0721	.1334

Table 7: Components in the product title "Women's Down Jacket Sleeve Covers, White Arm Warmers, Cotton Office Work Sleeves, Dirt and Stain-Proof, Arm Guards with Cuffs"

Component Type	Component Phrase
Central Category	"Sleeve Covers"
Synonym of Category	"Arm Warmers"
Description of features	"Women's Down Jacket", "White", "Cotton"
Highlight	"Dirt and Stain-Proof", "Arm Guards with Cuffs"
Application Scenario	" Office Work"

Table 7. In the product title, all the other phrases are related to the central category phrase "Sleeve Covers". For a BERT-based encoder, the attention distribution in the title is expected to highly focus on the central category phrase. There are almost no contextualized semantic relations among the other phrases (*e.g.*, Office Work" and "White"). This situation is completely different from a sentence in a passage. In conclusion, the product title is not a meaningful sentence. As a result, the sentence-based BERT encoders make it impossible to generate high-quality embeddings of products.

Sadly, since previous researchers have struggled with a lack of training data, most of those existing diversification approaches in web search are based on document representations [23]. As a result, those existing supervised diversification approaches cannot be adapted into diversification in product search directly. A possible way to measure the diversity of products is to model the fine-grained token-level interactions between different products [23]. Another possible solution is to design a contrastive learning approach especially optimized for the representations of products [9]. Since the amount of training data in JDivPS is significantly larger than the TREC WebTrack 2009-2012 datasets, it's possible to train those models mentioned above based on the pre-trained BERT model we provided alongside the dataset.

In a nutshell, these overall results demonstrate that it is necessary and valuable to design diversification approaches specific for product search with our JDivPS dataset.

5.5 Dataset Ablation Studies

In this section, we conduct ablation studies on MMR with different relevance models and product characteristics to further explore the impact of these features on the performance of diversified product search models. The variants include:

- MMR_{REL_name}, MMR_{REL_brand}, and MMR_{REL_cate}: We leverage the non-diversified model REL to generate the relevance score and the representation of product name, brand name and category name as the diversity features, respectively.
- MMR_{UVCTR_name}, MMR_{UVCTR_brand}, and MMR_{UVCTR_cate}: The non-diversified model UVCTR is utilized to produce the relevance score while the representation of product name, brand name and category name are treated as the diversity features, respectively.
- MMR_{EMB_name}, MMR_{EMB_brand}, and MMR_{EMB_cate}: We leverage the embedding of product name, brand name, and category name for both relevance and diversity calculation, respectively.

For results in Table 6, we can observe that (1) The diversified product search model significantly benefits from a better relevance calculation model. Concretely, the performance of models with UVCTR as the relevance model can outperform the other two with REL and EMB. (2) Leveraging brand names to calculate the differences between products can achieve the best results with any relevance score, which is consistent with the conclusion of our data analysis in Section 4 that brands have the highest proportion in

user intents. Besides, all those MMR-based approaches can only gain weak improvements compared with the baseline approaches. These results are identical to the analysis in Section 5.4.2, indicating that a carefully designed product representation model is necessary for the diversification of product search.

6 DISCUSSION OF APPLICATION SCENARIOS

To the best of our knowledge, our proposed JDivPS dataset is the first human-labeled dataset for the diversification of product search. As we described in Section 1, there are multiple existing works for the diversification of web search, but the diversification of product search is far less explored compared with web search. This is mainly because there are no publicly available datasets with real engaged queries and human-labeled user intents. Our dataset makes it available to propose more approaches for product search diversification. Here are the potential approaches we imagine based on our proposed dataset: first of all, as we mentioned in Section 5.4.2, a critical issue is that the existing approaches of sentence representation cannot be directly adapted to generate high-quality product representations. A possible solution is to propose a representation approach specifically designed for product representations. Besides, compared with web pages which are usually extracted into flattened tests, the product items preserve both names and other structural metadata, which can be processed separately by the ranking models.

Concatenating all those metadata phrases into flattened texts, our dataset can also be used for web search diversification. The existing WebTrack datasets involve only 198 queries with intent annotations, this data amount is extremely difficult to support ranking models based on token-level fine-grained document interactions. Our proposed JDivPS dataset is about 20 times larger than the WebTrack dataset, indicating that the researchers can easily apply ranking models on a larger scale to achieve better results. Although our dataset is anonymized, the research directions mentioned above may not be affected hardly since we provide a pre-trained BERT encoder alongside the dataset.

7 CONCLUSION

In this work, we introduce a new dataset JDivPS for diversification in product search. Compared with previous non-public and lacking human-labeled user intents datasets, our dataset is public and includes both real queries and human-annotated real intents. The statistical features of products in the dataset are also provided to demonstrate the products' popularity and commercial value. These bring our proposed JDivPS dataset closer to the distribution of actual user intents. Further investigation of product search diversification is made feasible by this dataset.

ACKNOWLEDGMENTS

Zhicheng Dou is the corresponding author. This work is done during Zhirui Deng's internship at JD.com Inc. This work was supported by the National Natural Science Foundation of China (62272467), and Public Computing Cloud, Renmin University of China. The work was partially done at the Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE, and Beijing Key Laboratory of Big Data Management and Analysis Methods.

REFERENCES

[1] 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, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 379–388. https://doi.org/10.1145/3357384.3357980

[2] Qingyao Ai, Yongfeng Zhang, Keping Bi, Xu Chen, and W. Bruce Croft. 2017. Learning a Hierarchical Embedding Model for Personalized Product Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 645–654. https://doi.org/10.1145/3077136.3080813

[3] Keping Bi, Qingyao Ai, and W. Bruce Croft. 2020. A Transformer-based Embedding Model for Personalized Product Search. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 1521–1524. https://doi.org/10.1145/3397271.3401192

[4] Jaime G. Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel (Eds.). ACM, 335–336. https://doi.org/10.1145/290941.291025

[5] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy Lin (Eds.). ACM, 621–630. https://doi.org/10.1145/1645953.1646033

[6] Xiangru Chen, Haofen Wang, Xinruo Sun, Junfeng Pan, and Yong Yu. 2011. Diversifying product search results. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft (Eds.). ACM, 1093–1094. https://doi.org/10.1145/2009916.2010065

[7] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong (Eds.). ACM, 659–666. https://doi.org/10.1145/1390334.1390446

[8] Charles L. A. Clarke, Maheedhar Kolla, and Olga Vechtomova. 2009. An Effectiveness Measure for Ambiguous and Underspecified Queries. In *Advances in Information Retrieval Theory, Second International Conference on the Theory of Information Retrieval, ICTIR 2009, Cambridge, UK, September 10-12, 2009, Proceedings (Lecture Notes in Computer Science, Vol. 5766)*, Leif Azzopardi, Gabriella Kazai, Stephen E. Robertson, Stefan M. Rüger, Milad Shokouhi, Dawei Song, and Emine Yilmaz (Eds.). Springer, 188–199. https://doi.org/10.1007/978-3-642-04417-5_17

[9] Zhirui Deng, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2024. CL4DIV: A Contrastive Learning Framework for Search Result Diversification. ACM. https://doi.org/10.1145/3616855.3635851

[10] 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*. 581–590.

[11] Kohei Hirata, Daichi Amagata, Sumio Fujita, and Takahiro Hara. 2022. Solving Diversity-Aware Maximum Inner Product Search Efficiently and Effectively. In *RecSys '22: Sixteenth ACM Conference on Recommender Systems, Seattle, WA, USA, September 18 - 23, 2022*, Jennifer Golbeck, F. Maxwell Harper, Vanessa Murdock, Michael D. Ekstrand, Bracha Shapira, Justin Basilico, Keld T. Lundgaard, and Even Oldridge (Eds.). ACM, 198–207. https://doi.org/10.1145/3523227.3546779

[12] Kohei Hirata, Daichi Amagata, Sumio Fujita, and Takahiro Hara. 2023. Categorical Diversity-Aware Inner Product Search. *IEEE Access* 11 (2023), 2586–2596. https://doi.org/10.1109/ACCESS.2023.3234072

[13] Kohei Hirata, Daichi Amagata, Sumio Fujita, and Takahiro Hara. 2023. Categorical Diversity-Aware Inner Product Search. *IEEE Access* 11 (2023), 2586–2596. https://doi.org/10.1109/ACCESS.2023.3234072

[14] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manag.* 36, 2 (2000), 207–227. https://doi.org/10.1016/S0306-4573(99)00056-4

[15] Yunjiang Jiang, Yue Shang, Ziyang Liu, Hongwei Shen, Yun Xiao, Sulong Xu, Wei Xiong, Weipeng Yan, and Di Jin. 2020. BERT2DNN: BERT distillation with massive unlabeled data for online e-commerce search. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 212–221.

[16] Zhengbao Jiang, Ji-Rong Wen, Zhicheng Dou, Wayne Xin Zhao, Jian-Yun Nie, and Ming Yue. 2017. Learning to Diversify Search Results via Subtopic Attention.

In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 545–554. https://doi.org/10.1145/3077136.3080805

[17] Jiongnan Liu, Zhicheng Dou, Guoyu Tang, and Sulong Xu. 2023. JDsearch: A Personalized Product Search Dataset with Real Queries and Full Interactions. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Taipei</city>, <country>Taiwan</country>, </conf-loc>) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2945–2952. https://doi.org/10.1145/3539618.3591900

[18] Jiongnan Liu, Zhicheng Dou, Qiannan Zhu, and Ji-Rong Wen. 2022. A Category-aware Multi-interest Model for Personalized Product Search. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini (Eds.). ACM, 360–368. https://doi.org/10.1145/3485447.3511964

[19] Xusheng Luo, Luxin Liu, Yonghua Yang, Le Bo, Yuanpeng Cao, Jinghang Wu, Qiang Li, Keping Yang, and Kenny Q. Zhu. 2020. AliCoCo: Alibaba E-Commerce Cognitive Concept Net. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (Portland, OR, USA) *(SIGMOD '20)*. Association for Computing Machinery, New York, NY, USA, 313–327. https://doi.org/10.1145/3318464.3386132

[20] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.

[21] Nish Parikh and Neel Sundaresan. 2011. Beyond relevance in marketplace search. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, Craig Macdonald, Iadh Ounis, and Ian Ruthven (Eds.). ACM, 2109–2112. https://doi.org/10.1145/2063576.2063902

[22] Xubo Qin, Zhicheng Dou, and Ji-Rong Wen. 2020. Diversifying Search Results using Self-Attention Network. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 1265–1274. https://doi.org/10.1145/3340531.3411914

[23] Xubo Qin, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2021. Interaction-Based Document Matching for Implicit Search Result Diversification. In *Information Retrieval*, Hongfei Lin, Min Zhang, and Liang Pang (Eds.). Springer International Publishing, Cham, 3–15.

[24] Craig Silverstein, Monika Rauch Henzinger, Hannes Marais, and Michael Moricz. 1999. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum* 33, 1 (1999), 6–12. https://doi.org/10.1145/331403.331405

[25] Ruihua Song, Zhenxiao Luo, Ji-Rong Wen, Yong Yu, and Hsiao-Wuen Hon. 2007. Identifying ambiguous queries in web search. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy (Eds.). ACM, 1169–1170. https://doi.org/10.1145/1242572.1242749

[26] Zhan Su, Zhicheng Dou, Yutao Zhu, Xubo Qin, and Ji-Rong Wen. 2021. Modeling Intent Graph for Search Result Diversification. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 736–746. https://doi.org/10.1145/3404835.3462872

[27] Le Yan, Zhen Qin, Rama Kumar Pasumarthi, Xuanhui Wang, and Michael Bendersky. 2021. Diversification-Aware Learning to Rank using Distributed Representation. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 127–136. https://doi.org/10.1145/3442381.3449831

[28] ChengXiang Zhai, William W. Cohen, and John D. Lafferty. 2015. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. *SIGIR Forum* 49, 1 (2015), 2–9. https://doi.org/10.1145/2795403.2795405

[29] Chenyu Zhao, Yunjiang Jiang, Yiming Qiu, Han Zhang, and Wen-Yun Yang. 2023. Differentiable Retrieval Augmentation via Generative Language Modeling for E-commerce Query Intent Classification. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4445–4449.

[30] Guorui Zhou, Chengru Song, Xiaoqiang Zhu, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep Interest Network for Click-Through Rate Prediction. arXiv:1706.06978 [stat.ML]