sQuIrRel: Large-Scale Evaluation of E-commerce Query Classification Models

Anna Tigunova
tigunova@amazon.de
Amazon
Berlin, Germany

Ghadir Eraisha
eraishag@amazon.lu
Amazon
Luxembourg

Ezgi Akcora
akcora@amazon.lu
Amazon
Luxembourg

Abstract

Query-to-Product Type (Q2PT) is a crucial e-commerce query understanding signal, which directly influences search results relevance and customer UX experience. This imposes high standards on the industrial Q2PT classification models, which have to be regularly monitored for quality among all predicted product types and use cases at scale.

Existing solutions for such Q2PT model evaluation involve human-labeled datasets, which are usually small-scale and are costly to collect and refresh. Moreover, it is unrealistic to create ample human annotations for all e-commerce product categories, which can span several thousands.

To address these drawbacks, we propose a method sQuIrRel (Query Intent from Relevance) to *automatically* collect an evaluation dataset for monitoring e-commerce query classification models, which ensures large-scale analysis and full coverage of all existing category labels. sQuIrRel is constructed using distant supervision from a high-precision query-item relevance classifier, allowing to quickly collect and refresh query labels at scale.

While sQuIrRel method can be applied to any query classification task, across various e-commerce stores, our study focuses on using sQuIrRel for Q2PT prediction. We provide comparisons with alternative dataset collection methods and show how the obtained dataset can be used to analyze the performance of a commercial Q2PT model.

CCS Concepts

• **Information systems** → **Query intent**.

Keywords

Electronic commerce, Model Evaluation Dataset, Search Query Classification

ACM Reference Format:
Anna Tigunova, Ghadir Eraisha, and Ezgi Akcora. 2025. sQuIrRel: Large-Scale Evaluation of E-commerce Query Classification Models. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3701716.3715261

WWW Companion '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/25/04
https://doi.org/10.1145/3701716.3715261

1 Introduction

E-commerce query understanding extracts customer intent from their search queries, predicting brand, color or target audience of the desired product. These signals help to refine search results and surface the most relevant products.

One of the most important query understanding signals is Query-to-Product Type (Q2PT), associating search keywords to the fine-grained product categories, such as '*toys and games*' or '*downloadable movies*'. Q2PT predictions are used in multiple e-commerce downstream components, reducing retrieval latency and improving customers' search and navigation experiences.

Q2PT is a very challenging task because of an extremely large label space and highly imbalanced label distribution. Moreover, international marketplaces face additional challenges such as multilinguality and locale-specific differences [5]. Given the importance of the signal and the difficulty of predicting it, industrial Q2PT models are required to be extremely reliable- and highly performant. This involves regularly monitoring the models at scale, across various languages and use cases.

Q2PT models are normally trained and validated using aggregated click-through data [2, 3, 8, 9]. This approach is, however, inadequate for reliable evaluation of the customer-facing model, as the click data is noisy and prone to exposure bias.

Alternatively, e-commerce companies carry out evaluation of Q2PT models using high-quality human-labeled datasets [6, 7]. This method, however, has numerous shortcomings: i) collecting human annotations is an extremely laborious, long and expensive process; in particular, it does not allow for frequent dataset refreshes, which are essential in a constantly evolving e-commerce domains; ii) in real e-commerce services there are thousands of product types with an extremely skewed distribution; as the queries for human annotation are usually sampled randomly, the resulting dataset will likely not cover all long-tail product types with an ample number of samples.

In this work we address the issues of both approaches by proposing an *automatically labeled* Q2PT evaluation dataset, dubbed sQuIrRel (Query Intent from Relevance).

Our dataset is labeled with distant supervision from the query-item relevance signal: the correct query product type is inferred from the catalogue items, which are an exact match for this query. The query-item matching score is predicted by a highly accurate bi-encoder model, fine-tuned on the relevance task. By this means, we completely automate large-scale collection of Q2PT data, guaranteeing high sample coverage of all existing product types. Additionally, this approach allows to automatically refresh the dataset, in case of the input/label distribution shifts.

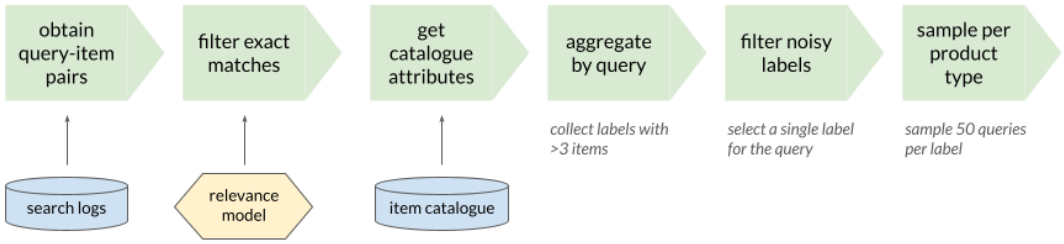

Figure 1: sQuIrRel dataset construction steps.

Note, that in sQuIrRel we directly leverage store-specific knowledge, coming from the item catalogue and query-items relevance signal, which would be impossible with the general knowledge of out-of-the-box LLMs.

Putting sQuIrRel in practice, we collected a dataset with 2.7M samples across 20 locales and over 1.5k product types. We manually verified the quality of the labels in sQuIrRel, obtaining over 87% accuracy.

We evaluate the utility of sQuIrRel by conducting an in-depth analysis of an existing Q2PT classifier performance. This allows us to identify gaps in the model's performance, such as: i) inability to handle fine-grained media product types (such as '*movie*' vs '*tv-season*'), ii) confusion between semantically close product types (e.g '*shirt*' vs '*tunic*'). Such analysis allows to identify avenues for the Q2PT signal improvement to ensure positive customer experience.

2 Methodology

In this section we first define the desired requirements of the Q2PT evaluation dataset, then we describe the pretrained relevance classifier used for weak supervision, and finally we outline the stepwise data collection procedure.

2.1 Overview

The evaluation dataset for Q2PT models needs to adhere to the following requirements:

(1) **High precision of annotations.** The evaluation data needs to reliably assess the model's performance, considering that a lot of downstream systems depend on the signal's quality.
(2) **Low cost to (re-)create.** On e-commerce sites the selection of items and the customer preferences are constantly evolving, causing changes in product type label space. Thus, the evaluation dataset needs to be easily refreshed or, in case of new country expansion, promptly created from scratch.
(3) **High coverage of existing product types.** In large e-commerce stores there are thousands of product types, most of which are long-tailed (e.g. '*snowblower*'). The evaluation dataset needs to cover all product types, to enable fine-grained Q2PT model evaluation.
(4) **Incorporate store-specific knowledge.** The evaluation dataset needs to reflect the natural customer queries and their real shopping intents, which requires marketplace-related knowledge.

One common approach to automatically collect query classification data is to utilize historical click-through data as a source of supervision. In this approach the query label is inferred from the items that users clicked following their search, under the hypothesis that the customers would click on items aligned with the intended product type. This method has been widely used in related works [3, 8, 9] for training and evaluation of query classification models. However, this approach does not adhere to the requirement (1), as user click behavior is noisy and unreliable, suffering from exposure bias, trends and seasonality.

An alternative method, which ensures high annotation quality is to use human judgement to create the evaluation data [7]. Yet, this does not fulfill requirements (2) and (3). Besides, human annotation is extremely time-consuming, which results in delayed feature launches and stale evaluation datasets.

Finally, one might opt for using an LLM's world knowledge to directly generate evaluation examples for a given product type. However, out-of-the-box LLMs are not aware of the peculiarities of a real e-commerce site, including user behavior and catalogue information, e.g. brand availability; this violates the requirement (4). Moreover, in our pilot experiments, we noticed that LLMs tend to generate unnatural queries, lacking conciseness and conversational features, such as the use of abbreviations.

We propose *sQuIrRel* (Query Intent from Relevance), which overcomes the drawbacks of the alternative methods and fulfills all outlined requirements.

Similar to the click-through aggregation approach, in sQuIrRel we infer query labels from the associated items. However, instead of relying on the noisy customer behavior to select these items, we propose to use supervision from a pretrained relevance model to identify the relevant items for each query. The label for the query will be then derived from *exact match* items, predicted by the relevance model. Our proposed approach overcomes the randomness of user clicks by leveraging the knowledge of the pretrained model.

2.2 Relevance model

The key component in our approach is the pretrained relevance model, assigning ESCI labels (*exact*, *substitute*, *complement*, *irrelevant*) to the query-item pairs. We used a multilingual BERT-based classifier, fine-tuned on human-labeled query-item pairs, sampled from real e-commerce data.

The fine-tuned relevance model achieved over 0.9 precision in predicting the *exact* label, which is used to select relevant items in sQuIrRel. Thus we ensure that the labeling supervision comes from a source that accurately captures the customer shopping intent.

2.3 Dataset construction process

In this work we leverage internal data for our dataset collection procedure. However, sQuIrRel can be used in any other e-commerce store with minimal modifications.

We made a design decision to exclude from sQuIrRel queries with a broad intent, which yield multiple product types (e.g., 'gifts for kids'). This choice is motivated by the fact that it is impossible to derive a single correct set of labels for these ambiguous queries, and thus to objectively assess the Q2PT model performance.

The dataset construction process consists of the following steps:

(1) **Obtain query-item pairs.** To maximize the chances of obtaining exact match query-items pairs, we use e-commerce search logs, containing queries and the list of all products that are returned to the customer.
(2) **Filter exact matches.** Each query-item pair is labeled with the relevance classifier, described in Section 2.2; we retain only the *exact* match pairs, according to the classifier. We additionally refine the exact match restriction, by discarding all pairs with classifiers' confidence score lower than 0.8.
(3) **Get catalogue attributes.** We use product catalogue to obtain the product types of all items in each query-item pair.
(4) **Aggregate by query.** For each query we collect the product types from its exact match items, discarding the product types that are supported by less than 3 items per query.
(5) **Filter noisy labels.** For each query we remove all product types that have less than 1/3 of exact match items. This process eliminates tail product types that are only moderately related to the query, and queries whose product types are too dispersed (broad intent queries). After this filtering, the number of possible product types per query will be in the range 0..2. As a final step, we select only queries with exactly one product type, as we do not address the case for the multi-PT queries.
(6) **Sample per product type.** We aggregate the queries per product type and we randomly downsample the dataset, to have at most 50 queries for each product type.

We sample around 1.5k product types and 20 countries to construct an experimental dataset for our analysis. We run sQuIrRel on 20 locales separately, as the catalogue availability and product type attribution can vary for different locales [5].

3 Dataset analysis

The resulting experimental dataset contains over 2M of queries, with around 130k queries per locale on average.

We observed that in some locales the dataset lacks specific product types; this observation does not mean the deficiency of sQuIrRel, but rather it reflects the state of the marketplace in individual countries. Such missing product types fall into specific categories, such as fresh groceries (not sold in every locale), restricted items (alcohol, weapons), or digital content. In some cases the missing product types simply do not make sense to be sold in the given country (e.g., snowboarding equipment or snow cleaning devices in the middle east countries).

Manual verification. To validate the quality of sQuIrRel we ran a manual verification of the resulting Q2PT labels. We sampled over 400 random labeled queries from 8 locales (Egypt, Netherlands, Singapore, Germany, Turkey, Italy, Mexico, United States) and asked individual annotators, proficient in the respective language, to evaluate the labeling.

locale	query	click-through	sQuIrRel
SG	integrated graphics	NOTEBOOK_COMPUTER	VIDEO_CARD
SG	mop bucket	MOP_BUCKET_SET	BUCKET
UK	carry on luggage	BACKPACK	SUITCASE
UK	outdoor cushions	DECORATIVE_PILLOW_COVER	PILLOW

Table 1: Examples of different labels obtained from sQuIrRel and from aggregated click data.

This verification resulted in 87% accuracy of the collected dataset. The only systematic error, outlined by the annotators, was a significant number (around 20%) of complimentary product type confusions (e.g., assigning *'vacuum cleaner'* product type for the query 'vacuum cleaner handle').

3.1 Comparison with aggregated click data

We evaluate how sQuIrRel compares to the popular approach of collecting query understanding labels from click-through logs [3, 8, 9]. For that we have sampled a hundred queries from our dataset for Singapore (small store) and United Kingdom (high-resource store) and derived the labels for them from the aggregated click data.

In this method the product type for the search query is derived as the majority product type among the items, that users clicked following that query. Formally, we define the probability of the query q having the product type p as follows:

$$P(q,p) = \frac{num\_clicks(q, item|item \in p)}{num\_clicks(q, item)} \tag{1}$$

For our click-through dataset we select all <*query, product type*> pairs that have $P(q,p) > 0.5$.

We observed that sQuIrRel and click-through annotations yield similar results for many queries, but their differences are amplified in the smaller store (64% alignment for Singapore, 81% for the UK). We manually examined the cases where the predictions differ, some examples are presented in Table 1. These differences illustrate that using the noisy click data results in imprecise labels.

One reason for the inaccurate click-though labels is the mismatch between the users' browsing intent and the literal meaning of their query. For example, the query 'integrated graphics' literally means a *'video card'*, but the customers issuing this query mostly click on the *'notebooks'* that have this particular video card inside.

Another source of the click distribution inaccuracy is the customers changing their purchase intent after examining the search results (e.g., opting for a cheaper product type). For instance, the query 'carry on luggage', the product type *'suitcase'* predicted by sQuIrRel sounds more applicable; however, the customers preferred to purchase a cheaper and more practical *'backpack'*, which is technically a correct product type for this query too. Similarly, for the query 'outdoor cushions', the users instead of buying a whole new *'pillow'* (which is correctly labeled in sQuIrRel), decided to buy a cheaper *'pillow cover'* to upgrade their existing pillow.

top performing PTs	least performing PTs
CELLULAR_PHONE_CASE	DIP_SWITCH
INTERNAL_MEMORY	TV_SEASON
COMPUTER_DRIVE_OR_STORAGE	MUSIC_TRACK
KEYBOARDS	SOFTWARE
TELEVISION	MUSIC_ALBUM

Table 2: Some of the best and least performing product types by recall at 0.8 precision.

3.2 Comparison with human-labeled data

We compare sQuIrRel with the in-house manually annotated dataset, previously collected to evaluate the existing query understanding models. Although human-labeled data is a preferred source for model evaluation in many domains, there are several salient disadvantages of it for Q2PT evaluation:

- **Size and coverage.** Obtaining a large amount of manual labels is prohibitively expensive even in industrial settings. The human-labeled dataset that we consider is about 30 times smaller than sQuIrRel. Importantly, out of over 1k product types, only 600 were present in the manual dataset, with most of them containing an insufficient amount of samples.
- **Freshness.** The human-labelled dataset cannot be easily adjusted to the changes in label space and customer preferences, requiring being recomputed from scratch.
- **Quality.** In many cases assigning a product type to the query is a subjective task and it requires some domain expertise (e.g., the knowledge of brand and model names). Moreover, when deciding on a label, it is complicated for human annotators to take into consideration several hundreds of existing product types and their descriptions.

However, an advantage of the human-labeled data is that annotators can assign multiple labels to a query, while sQuIrRel is designed to produce only a single label.

For our comparison we selected a subset of queries that overlap between human-annotated dataset and sQuIrRel; we categorize the observed differences into 3 classes:

1. **Synonymous assignments:** for some queries several product types can be used interchangeably. For example, the query 'biscuits' was classified as a *'cracker'* by humans and *'cookie'* in sQuIrRel, while both labels are applicable.
2. **Disambiguation issues:** some queries allow several interpretations because the customer intent is unclear. For example, in Singapore store the query 'follow me' was classified as *'shampoo'* by humans, which is a valid Malaysian brand, returned among the top results for this query. At the same time, sQuIrRel labeled this query as *'book'*, because of the large number of books in the catalogue with 'follow me' words in the title.
3. **Multi-PT annotations**, related to the mentioned issue of sQuIrRel returning only a single product type; in such cases sQuIrRel labels will be a subset of the manually assigned.

4 Evaluating Q2PT models with sQuIrRel

In this section we demonstrate how the constructed dataset can be used for a fine-grained evaluation of an industrial Q2PT classifier. sQuIrRel contains sufficient number of samples for each existing product type, which allows us to report per-product type results.

We train a BERT-based Q2PT classifier for all 20 marketplaces, following [5]. The model is trained using large-scale aggregated click-through data using Binary Cross-Entropy loss.

To evaluate the resulting classifier, we compute per-PT recall at 0.8 precision. The choice of this metric is motivated by the requirement to have high-precision query understanding models for the downstream applications. Per-PT recall is defined as the number of correctly predicted occurrences of a product type, over all occurrences of this product type in the ground truth data. For comparison, the macro-averaged recall of the overall dataset is 0.82.

In Table 2 we show some of the best and the least performing product types in terms of recall. We can observe that the model does well on some electronics-related product types (e.g. *'internal memory'* or *'keyboards'*), while it struggles to make correct predictions for some media product types (like music content and software).

In Table 3 we show some of the instances where the model demonstrated confusion among product types. We classify the model's confusions into the following categories: i) confusing media-related product types to *'books'*, because books is prevalent among media items, having similar titles to movies and series, ii) confusing synonymous or interchangeable product types (such as *'tunic'* and *'shirt'*), and iii) confusing complimentary product types (e.g. *'bed'* and *'bed frame'*).

Such analysis is crucial for industry practitioners, helping to identify the query understanding model's defects and discover avenues for signal improvement, ensuring an optimal user experience.

5 Related work

Most studies on query classification use aggregated click-through data for both model training and validation [1–3, 8, 9], mitigating the noise by setting thresholds on <*query, product type*> click scores. A twist on that is proposed by Zhang et al. [6], who introduce an approach to augment the incomplete click-though NER annotations by distilling the scores from the model, trained on the human-labelled dataset. Shen et al. [4] investigate training a Q2PT model on product titles and on pseudo-queries constructed from the product titles, but find that the model trained on click-through data still performs better. Finally, only few studies use human-labeled datasets for evaluation [6, 7], due to the cumbersome and costly manual annotation.

PT	confused to
TUNIC	SHIRT
TV_SEASON	BOOK
BED	BED_FRAME
MOVIE	BOOK
SKILL_APPLICATION	BOOK
MUSIC_TRACK	BOOK
AIR_GUN	TOY_GUN
FLAT_SHEET	BED_LINEN_SET
MODEM	NETWORKING_ROUTER

Table 3: Some of the confused product types.

6 Conclusion

In this work we introduce sQuIrRel, an automated procedure to construct large-scale evaluation data for query understanding models, guided by relevance supervision.

We instantiate our method on query-to-product type prediction, which is a crucial signal used in many internal components in e-commerce. The proposed approach alleviates the cost of human annotation and accelerates experimentation. sQuIrRel method is highly scalable, it can be easily expanded to further query understanding signals, specific usecases and different e-commerce stores.

As future work, we plan to experiment with applying sQuIrRel to other query understanding attributes, such as brand classification. We also aim to address labeling of broad and vague queries. Finally, we plan to explore mixing click-through data with the relevance signals, to get more holistic representation of the customer intent.

References

[1] Yiu-Chang Lin, Ankur Datta, and Giuseppe Di Fabbrizio. 2018. E-commerce product query classification using implicit user's feedback from clicks. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 1955–1959.

[2] Xianjing Liu, Hejia Zhang, Mingkuan Liu, and Alan Lu. 2019. System Design of Extreme Multi-label Query Classification using a Hybrid Model.. In *eCOM@ SIGIR*.

[3] Yiming Qiu, Chenyu Zhao, Han Zhang, Jingwei Zhuo, Tianhao Li, Xiaowei Zhang, Songlin Wang, Sulong Xu, Bo Long, and Wen-Yun Yang. 2022. Pre-training tasks for user intent detection and embedding retrieval in e-commerce search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4424–4428.

[4] Dou Shen, Ying Li, Xiao Li, and Dengyong Zhou. 2009. Product query classification. In *Proceedings of the 18th ACM conference on information and knowledge management*. 741–750.

[5] Anna Tigunova, Thomas Ricatte, and Ghadir Eraisha. 2024. Transfer Learning for E-commerce Query Product Type Prediction. *arXiv preprint arXiv:2410.07121* (2024).

[6] Danqing Zhang, Zheng Li, Tianyu Cao, Chen Luo, Tony Wu, Hanqing Lu, Yiwei Song, Bing Yin, Tuo Zhao, and Qiang Yang. 2021. Queaco: Borrowing treasures from weakly-labeled behavior data for query attribute value extraction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4362–4372.

[7] Junhao Zhang, Weidi Xu, Jianhui Ji, Xi Chen, Hongbo Deng, and Keping Yang. 2021. Modeling Across-Context Attention For Long-Tail Query Classification in E-commerce. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 58–66.

[8] Lvxing Zhu, Hao Chen, Chao Wei, and Weiru Zhang. 2022. Enhanced representation with contrastive loss for long-tail query classification in e-commerce. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*. 141–150.

[9] Lvxing Zhu, Kexin Zhang, Hao Chen, Chao Wei, Weiru Zhang, Haihong Tang, and Xiu Li. 2023. HCL4QC: Incorporating Hierarchical Category Structures Into Contrastive Learning for E-commerce Query Classification. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 3647–3656.