EasyChair Preprint

№ 7347

WANDS: Dataset for Product Search Relevance Assessment

Yan Chen, Shujian Liu, Zheng Liu, Weiyi Sun, Linas Baltrunas and Benjamin Schroeder

January 21, 2022

WANDS: Dataset for Product Search Relevance Assessment

Yan Chen, Shujian Liu, Zheng Liu,
Weiyi Sun, Linas Baltrunas, Benjamin Schroeder

{ychen4, sliu1, zliu2, wsun1, lbaltrunas, beschroeder}@wayfair.com
Search and Recommendation, Wayfair, U.S.A

Abstract. Search relevance is an important performance indicator used to evaluate search engines. It measures the relationship between users' queries and products returned in search results. E-commerce sites use search engines to help customers find relevant products among millions of options. The scale of the data makes it difficult to create relevance-focused evaluation datasets manually. As an alternative, user click logs are often mined to create datasets. However, such logs only capture a slice of user behavior in the production environment, and do not provide a complete set of candidates for annotation. To overcome these challenges, we propose a systematic and effective way to build a discriminative, reusable, and fair human-labeled dataset, Wayfair Annotation DataSet (WANDS), for e-commerce scenarios. Our proposal introduces an important cross-referencing step to the annotation process which significantly increases dataset completeness. Experimental results show that this process is effective in improving the scalability of human annotation efforts. We also show that the dataset is effective in evaluating and discriminating between different search models. As part of this contribution, we also released the dataset. To our knowledge, it is the biggest publicly available search relevance dataset in the e-commerce domain.

Keywords: product search · search Relevance · dataset · evaluation

1 Introduction

Search engines are a big part of our day-to-day lives. They are behind many applications we have come to rely on daily, from web retrieval to e-commerce. Thus, it is hardly a surprise that a lot of research has been poured into improving and evaluating search engines. Search relevance is a measure of the accuracy of the relationship between the search query and the search results. It is commonly used to assess the performance of search engines.

Evaluating search relevance is inherently tricky. It is a common practice to use annotators to indicate the relevancy of a query-result pair. However, on a large scale, it is not possible to ensure the completeness of the evaluation set. The purpose and use case of queries also vary significantly, which makes discerning the intent of the query a challenge. This, in turn, makes it hard to pinpoint the exact

search results that are expected. For example, if a user is interested in finding *induction cooktops*, and attempts to search for them using the query *cooktop*, it poses an interesting annotation challenge - how do we discern between results which include only induction cooktops from those which return all cooktops?

In this paper, we introduce and describe *WANDS*, an open-source e-commerce product dataset that can be used to fairly and accurately evaluate the relevancy of e-commerce product search engines. We will explain our data collection methodology, as well as share experiments that we have conducted to validate the efficacy and value of *WANDS*. The key contributions of this paper include:

- releasing a public dataset, which is built on top of real-world e-commerce production data. To the best of our knowledge, this is the biggest search relevance dataset in the e-commerce domain.
- detailing the methodology used to construct the dataset to allow for transparency and reproducibility.
- proposing an iterative product mining technique called "cross-referencing" to improve the completeness of our annotations while keeping the annotation problem tractable.

2 Related Work

There has been a sizable body of work created on the problem of evaluating search relevance. We partition this prior work into Web Search Relevance and Product Search Relevance.

Web Search Relevance deals with retrieving unstructured search responses from large web-scale datasets. The best-known body of work around web-scale relevance evaluation is from the Text REtrieval Conferences (TREC), a series of evaluation workshops conducted for several years. TREC 2007 and 2008 featured the million query track [6, 5] which involved searching over the `GOV2` dataset [2]. The dataset used is a collection of web pages from within the `.gov` domain, and includes around 25 million documents. Part of the track's goal was to investigate whether multiple shallow judgments might be a better alternative to using fewer, more thorough judgments. The 2009 run of this track [9] used a new `ClueWeb09` dataset [3] instead of `GOV2`. This is a much larger dataset of one billion web pages in 10 languages.

Besides academia, multiple enterprises in the tech industry have also shared their research in this space. Google released a sample of their internal annotation guidelines [1]. While it provides a useful peek at how they define relevance, the guidelines do not shed sufficient insights into what Google defines as a "*best*" match. Microsoft Bing made available a package of benchmark dataset *LETOR* [18] for learning to rank, which contains standard features,

[1] https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

data, and evaluation tools. Sogou, a Chinese search engine, released *SOGOU-SRR* (Sogou Search Result Relevance) [27] and *SOGOU-QCL* [29]. These are large and high-quality datasets. However, these datasets would not be the most appropriate for evaluating product search relevance, since their ranking target is web pages instead of products.

Product Search Relevance focuses on retrieving items from datasets of products and merchandise. The community has adopted two main approaches to build product search relevance datasets: mining user click logs and annotating via crowdsourcing.

Mining user click logs is a popular way to build up significantly sized datasets for large enterprises which have ready access to these logs. The rising popularity of embedding-based product retrieval [24, 15, 25, 26, 28] is facilitated by datasets assembled from these web-scale search logs. However, these datasets can be noisy as users can click on irrelevant but popular products, and also because non-clicks are difficult to interpret in terms of relevance. Moreover, such datasets are proprietary and have not been released to the public domain.

Datasets in the public domain tend to be crowdsourced datasets that do not leak proprietary and important data. The two following datasets are closely related to *WANDS*.

- Home Depot[2] released the "*Home Depot Product Search Relevance Dataset*" [1, 10] on Kaggle. It contains 75K training data samples and 166K evaluation samples. Queries are sampled from Home Depot's search logs. Ground truth labels, between 1 (not relevant) to 3 (highly relevant), are created via crowdsourcing. Each annotation was evaluated by at least three human raters, and the final relevance score is defined to be the average of these human ratings.
- Crowdflower also released a dataset [4] that contains relevance annotations from several e-commerce sites. This is a smaller dataset than the Home Depot one, including 261 search terms and a list of products for each of these terms. Annotations are based on a sliding scale from 1 to 4, where 4 indicates that the product fully satisfies the search query, and 1 indicates that the product does not match a query.

Both of these datasets only provide relevance scores or labels for their training samples, but not for testing samples. This reduces the usability of these datasets for benchmarks and comparison purposes.

Compared to the Home Depot and Crowdflower datasets, *WANDS* is significantly larger in terms of annotated query and product pairs. It includes relevance labels for both training and evaluation datasets to facilitate benchmarking and comparisons. Unlike the two existing datasets, with our *WANDS* dataset we will also release the full annotation guidelines we used, to ensure reproducibility and also to share best practices for future data collectors. *WANDS* also innovates on the annotation process to improve the number of relevant products per query (i.e., the cross-referencing process described in Section 4).

[2] Major U.S.A. home improvement retailor: http://www.homedepot.com

3 Annotation Guidelines Design

3.1 Design Principles

In this section, we will detail the design of our annotation guidelines. We design the *WANDS* dataset to meet the following criteria:

Reusable. Our dataset should apply to a wide variety of systems, and provide reproducible results. The most straightforward way to annotate an evaluation dataset is to present a particular Information Retrieval (IR) system's outputs to the annotators and to obtain human judgement specific to the IR system outputs. However, such annotation is not suitable for judging a different IR system. We aim to design a relevance dataset that can be used to evaluate multiple systems.

Fair [17]. It should be agnostic to the systems to be evaluated, and be able to evaluate product search engines fairly and objectively. As discussed in Section 2, user behavioral data becomes an increasingly popular choice as relevance evaluation datasets [24, 15, 25, 26, 28]. User behavior log data suffer from positional biases and would favor the rankings similar to the production system. We will alleviate positional bias issues by presenting pair-wise query and product information for annotators to judge.

Discriminative. It should have the power to discern the performance of different product search engines given a robust and discriminative evaluation metrics such as nDCG [20, 19, 14, 21]. In order to design a dataset that can differentiate great search algorithms from the good ones, we make sure to include hard negatives, the products are almost relevant to a query but not quite. We mined the hard negatives both arithmetically and from user behavior logs.

Completeness. As a core element in the Cranfield paradigm [11], completeness has been a debated quality of a relevance dataset since then [23]. Completeness refers to the property that within a relevance dataset, all relevant documents for a given query are known. Indeed, modern relevance datasets have mostly prioritized dataset size over completeness [22], and various evaluation metrics have been proposed to deal with incompleteness of evaluation datasets [7, 19]. However, as we will show in Section 6, incompleteness in Product Search dataset does negatively impact the discriminative power of the evaluation. Incompleteness in Product Search evaluation data also contributes to the problem that offline evaluation results cannot predict online metrics [13]. While we acknowledge that absolute completeness is impossible to achieve for a dataset the size of *WANDS*, we take measures to minimize the impact of incompleteness.

To understand why *completeness* is important, let's assume that we have a target query, which is expected to return 3 products $(p1, p2, p3)$ out of a set of 10. Let's assume that we have two versions of the dataset, A, and B. A includes 2 "relevant" annotations for $p1$ and $p2$. B includes 3 "relevant" annotations for $p1, p2, p3$. Suppose also that we have two search engines that we want to evaluate, α and β. α is able to return two results $(p1, p2)$, while β returns all three relevant products. When evaluated on dataset A, the two search engines will perform identically. It is not possible to tell them apart. However, we will

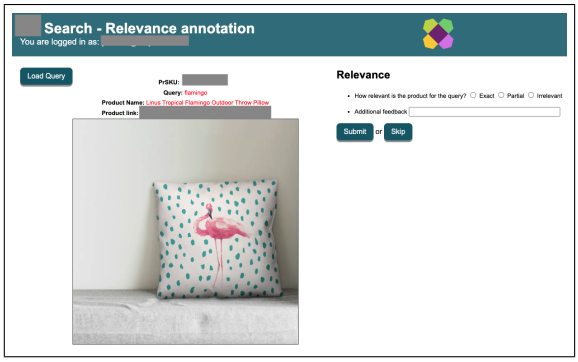

Fig. 1: Screenshot of Annotation Web UI.

be able to tell using dataset B that β is the better search engine, because it is able to return a better set of results than α.

3.2 Query-Product Annotation

We had three dedicated annotators from the data annotation agent working on this project. Each query, q_k, and its set of candidate products, $\theta_{q_k} = \{i_{k1}, i_{k2}, ..., i_{kn}\}$, are sent to the three annotators. The annotators see one query-product pair at a time and judge each query-product pairing with one of these possible annotations[3]:

- **Exact match:** The surfaced product fully matches the search query.
- **Partial match:** The surfaced product does not fully match the search query. It only matches the target entity of the query, but does not satisfy the modifiers for the query.
- **Irrelevant:** The product is not relevant to the query.

The annotators are given access to a web-based annotation tool to perform the labeling tasks as shown in Fig. 1.

4 Annotation Process

The overview of the annotation process is illustrated in Fig. 2. We started by stratified-sampling of search queries from a pool of historical customer queries stored in the e-commerce customer behavior logs. We then collected the products potentially relevant to one or more of the selected queries and constructed a Product Pool. Once the query and product pools were constructed, we performed Iterative Product Mining to identify the query-product pairs to be annotated. Three annotators then provided independent judgments on the selected query-product pairs, according to the Annotation Guidelines. To reduce dataset

[3] Please refer to our Annotation Guidelines released as a supplement to the dataset

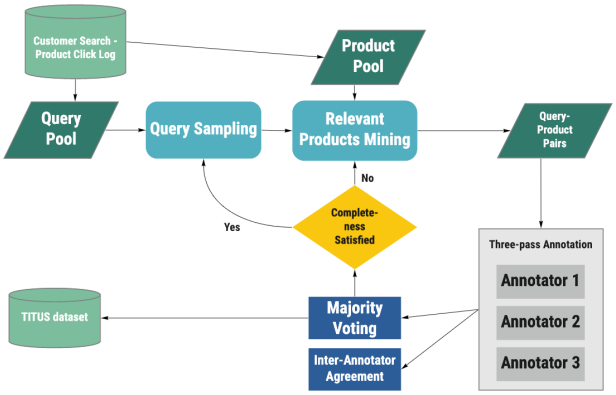

Fig. 2: Overview of the Annotation Process

incompleteness, we introduce the Iterative Product Mining process as described in Section 4.3. In the remainder of this section, we will discuss in detail each step of the annotation process.

4.1 Query Sampling

Our e-commerce website serves millions of queries every day. A good search relevance dataset should represent the diversity of real-world queries. To this end, we performed stratified sampling over tens of millions of customer search queries from the 2021 first quarter search log at our U.S. website. This resulted in a total of 480 English search queries.

Specifically, we stratified search queries along the following dimensions: 1) on-site organic searches vs. marketing-redirected searches, 2) searches that resulted in customer engagement (e.g., added products to cart) vs. searches that didn't result in customer engagement, and 3) popularity over the past two years. Within each stratified query group, we picked queries from both the head (frequent) and the tail (infrequent) of the frequency distribution. This approach improves the diversity of the queries in the query pool. Fig. 3 illustrated the diverse query distribution over the popularity and engagement dimensions.

4.2 Constructing the Product Pool

Our product catalog contains tens of millions of products. For this annotation task, we need to sample a small subset of our product catalog, such that the resulting relevance data set can differentiate great search models from good ones. This means that for the selected queries, we not only need to include clearly relevant products and clearly irrelevant products, but also need to ensure that there are hard-to-determine, almost-relevant products.

To mimic the real-world difficulty of a product search engine, we adopted two strategies to construct the product pool: using customer engagement data, and using a combination of lexical and neural retrieval systems:

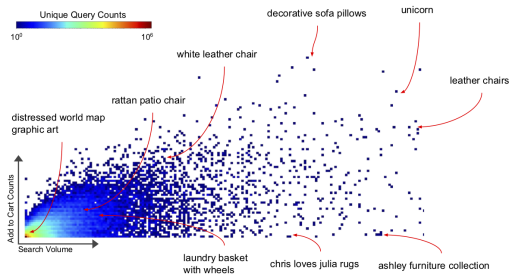

Fig. 3: Query Distributions on Add to Cart and Search Volume.

1. We leveraged user engagement data, and included the products that users clicked on or added to shopping cart during a search experience. Our hypothesis is that the user's added-to-cart products are good approximation of potentially relevant products[4], and their clicked-on (but not added-to-cart) products could be hard negatives, or almost-relevant products.
2. We further mine our product catalog using an open source lexical search engine Solr[5], and a neural product retrieval system inspired by Nigam et al. [16]. The two systems provide different ways to approximate relevant product retrieval. Neither system is perfect thus providing us more chances to include almost-relevant products in additional to relevant products.
3. We didn't attempt to sample easy negative samples (i.e. clearly irrelevant products). We assume that some of the selected products for certain queries will become negative samples for other queries.

4.3 Iterative Product Mining for Dataset Completeness

The query-product pairs resulting from the Product Pool Construction step were sent to the annotators as the first batch of annotation data. Pooling products related to different queries can cause dataset incompleteness [23]. In an ideal world, we would ask our annotators to judge every product and query pairs but that would be intractable - to do so in *WANDS* would require 60 million annotation judgements (480 queries x 42,994 products x 3 passes). To reduce the number of unjudged but relevant query-product pairs, we iteratively mined the entire product pool for unjudged but potentially relevant products for each query as cross-referencing. We presented the mined product-query pairs to the annotators in batches of decreasing likelihood of containing relevant pairs, and monitored the percentage of exact match query-product pairs in the annotation results. Once the percentage of exact match labels dropped to a predetermined

[4] Users purchasing irrelevant products in search results is a well documented phenomenon[8], however, it is not a concern in our case.

[5] https://solr.apache.org/

Table 1: Summaries of *WANDS* and other open-source datasets.

Feature	***WANDS***	**Home Depot**	**Crowdflower**
Query			
Counts	480	11,795	261
Predicted Class	✓	✗	✗
Product			
Counts	42,994	54,682	29,790
Primary Class	✓	✗	✗
Title	✓	✓	✓
Description	✓	✓	✓
Attributes	✓	✓	✗
Category Hierarchy	✓	✗	✗
Average Rating	✓	✗	✗
Number of Reviews	✓	✗	✗
Annotated Query-Product Relevance Labels			
Counts	233,448	74,067	22,513

level (5%), we would stop the product mining step and assume that the majority of relevant products had been found.

Specifically, we applied the lexical and neural retrieval systems described in Section 4.2 to discover more potentially relevant products. We further utilized a proprietary deep learning query classification model, which won during A/B test, to predict the product type that a certain query refers to (e.g. query "textured cotton throw pillow" was classified to "accent pillow" product class), and collected all the items in the product pool that belonged to this product class. After the iterative mining, we have reduced the chance of having unjudged but related query-product pairs in our dataset, and improved dataset completeness.

5 Dataset

The main contribution of this paper is the *WANDS* dataset[6] itself. We collected a total of 480 queries, 42,994 products, and 233K annotated query-product relevance labels. Table 1 shows a summary of *WANDS* relative to the Home Depot and Crowdflower datasets. *WANDS* contains the largest number of relevance labels for query-product pairs. It also contains the richest descriptions of the products and queries in the English language. It includes details such as: product title, product description, primary classes that product belongs to (i.e., chair), product category hierarchy, various product attributes such as size and color, average customer ratings, and review numbers.

Each entry in the dataset maps a (query-product) pair to a single relevance label, which could be one of 1) exact match, 2) partial match, or 3) irrelevant. This label is obtained by aggregating up to 3 entries from our annotators, using the majority vote strategy.

Quality Assurance is a common challenge for human-annotated datasets. Without a rigorous quality control strategy, annotators would produce an abundance of poor judgments. To ensure the quality of the annotations, we tracked

[6] https://github.com/wayfair/WANDS

Table 2: Change in inter-annotator agreement over time.

	Months 1 & 2	Month 3	Month 4
Cohen's Kappa	0.467	0.664	0.826
OPA	0.688	0.812	0.916

changes in inter-annotator agreement over time. We do this using two objective quality metrics: 1) Cohen's Kappa [12] and 2) the overlap percentage of agreement (OPA). Both metrics measure the agreement between raters, based on the judgments they make. OPA describes how frequently annotators agree with each other. For example, if 3 annotators all come to the same conclusion, then the inter-annotator agreement is 100%. If 2 out of 3 of them have the same conclusion, then the agreement is 66%. Table 2 shows the changes in inter-annotator agreement over a period of several months. The annotators started with moderate agreement, which steadily increased over the period of 4 months to an almost perfect agreement. Overall, there is a high level of agreement between our annotators leading to the high-quality dataset. We identified four reasons that contribute to significantly improved agreement: 1) Daily routine to discuss the conflicting annotations can help our annotators get calibrated to the annotation guidelines. 2) Regular audits and reviews help to train and align annotators. 3) With the input from annotators, we refine and fine-tune annotation guidelines. 4) Each query-product pair from a new annotator is also labeled by the other two annotators. This ensures data quality and facilitates future alignment. As annotators get trained and are more effective at the task, the overlapped examples are reduced to improve throughput.

Throughput is an important practical aspect of data collection. To determine the initial throughput, we piloted an annotation exercise with four team members. Following our annotation guidelines, we could achieve an initial throughput of 200 query-product pairs per hour, with an OPA of over 90%. The annotators performed consistently for the observed period of time. The throughput after 4 months is at around 190 query-product pairs per hour.

6 Experimental Evaluation

Datasets. Publicly available datasets are used to evaluate performance of search models. We prefer a dataset that provides statistically significant separation between competing search models. We designed an experiment to compare discriminative power of *WANDS* with other two public datasets. Home Depot [1] and Crowdflower [4] are public e-commerce product search datasets consisting of query and product pairs from popular e-commerce websites. Table 1 summarizes the differences between each of these datasets.

Search Models. For the experiment, we also needed to select a set of search models that by design, have known retrieval performance. Note that we are not evaluating the models themselves, but rather how well the dataset can differentiate between two similar models. We chose the following search models with known performance properties:

Table 3: Mapping of labels across each dataset to a standardized set of scores for metric computation.

Home Depot Dataset	Crowdflower Dataset	*WANDS*	Relevance Score
≥ 2.5	1	exact_match	1.0
≥ 1.5 and < 2.5	2	partial_match	0.5
	3		
≤ 1.5	4	irrelevant	0.0

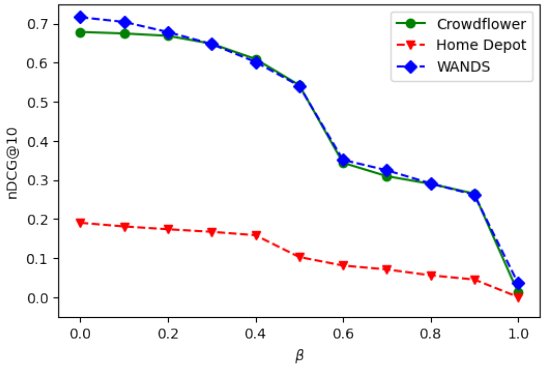

Fig. 4: nDCG@10 with varying β.

- **Random ranking (RANDOM)**. This is a naive baseline that generates a random list of products as the result to a query.
- **Okapi BM25 (BM25)**. That is a probabilistic retrieval ranking model which is integrated into `Apache Solr`. It is based on a bag-of-words representation and uses TF-IDF to estimate the relevance between query and products. This is a widely used and very popular unsupervised search algorithm. In our experiments, we make use of product titles and descriptions in each of the different dataset for ranking.
- **Linear combination of RANDOM and BM25 (LINEAR-β)**. This combines both RANDOM and BM25 linearly. The score of a product is computed as follows: $\beta * S_{RANDOM} + (1-\beta) * S_{BM25}$, where S_x denotes the score assigned to a product by system x and β is a parameter that defines mixing ratio between two base algorithms.

We use nDCG@10 [20] to evaluate the performance of different search models. To compute the metric, we have to resolve the differences in labels used across all three datasets. We map them to the relevance score as shown in Table 3.

Results. We ran *RANDOM*, *BM25* and *LINEAR-β* on three datasets, while varying the values of β from 0.0 to 1.0. Each experiment was repeated 5 times and values averaged. Note that performance of *LINEAR-0.0* is equivalent to *BM25* and *LINEAR-1.0* is equivalent to *RANDOM*. Figure 4 shows a plot of the nDCG@10 scores for varying levels of β on each of the experimental datasets.
Observations. The first observation is that nDCG@10 remains nearly constant across all values of β for the Home Depot dataset. This indicates the dataset is

Table 4: Label Distribution for different sampling sources.

	exact_match	**partial_match**	**irrelevant**
User click logs	8,420 (62.47%)	4,624 (34.31%)	434 (3.22%)
Open-source ranking systems	12,040 (30.95%)	24,797 (63.74%)	2,066 (5.31%)

not able to differentiate between search engines in terms of performance, even though *RANDOM* is expected to under-perform *BM25*. In fact, we could not differentiate between any LINEAR-β on Home Depot dataset (one-sided T-test, $p < 0.01$). On the other hand, on the Crowdflower and *WANDS* datasets, we can see an expected gradual decrease in nDCG@10 scores as the value of β increases. The graph is monotonically decreasing, with the highest nDCG@10 score for *BM25* (i.e. $\beta = 0.0$), and the lowest for *RANDOM* (i.e. $\beta = 1.0$). When comparing Crowdflower and *WANDS*, we can see that *WANDS* is more discriminative of the two. We can reject the null hypothesis that *LINEAR-0.0* and *LINEAR-0.3* have the same performance (one-sided T-test, $p < 0.01$). However, we cannot statistically separate *LINEAR-0.0* and *LINEAR-0.3* when using the Crowdflower dataset. For Crowdflower, we only see the same level of statistical significance for *LINEAR-0.0* and *LINEAR-0.5*. Thus, we conclude that WANDS has the highest discriminative power as compared to other datasets.

7 Discussion

7.1 Effectiveness of Sampling Sources

Constructing the product pool step uses two candidate sources: user click logs, and open-source ranking systems (e.g. `Solr`). Using user click logs is a popular way to gather query-product pairs and provide a valuable relevance signal. However, relying solely on click logs can lead to an incomplete dataset. This approach misses out on a lot of relevant items that users do not interact with. We augment this with results using open-source ranking systems. While these systems are imperfect, they do greatly expand the possible query-product pairings.

Table 4 shows the breakdown of the distribution of our annotation labels for each of these two sources. We see that the `exact_match` labels mined from user click logs are relatively high at 62.47%, and irrelevant candidates only account for 3.22% of all labels. For open-source ranking systems, we achieve around 31% of `exact_match` labels and 64% `partial_match` labels.

The high proportion of relevant matches from both approaches suggests that our sampling step is working as it was intended - to help narrow down a good list of candidates so that our annotators can pick out relevant matches efficiently.

7.2 Iterative Product Mining for Dataset Completeness Step

We analyzed iterative annotation process to understand how much value it added. After completing the step *Constructing the product pool*, for 49,390 query-product candidates and 480 queries, we obtained a set of 46,875 relevant (e.g.

Table 5: Distribution of labels across annotation steps 1 and 2.

	exact_match	**partial_match**	**irrelevant**
Step 1	18,018 (36.48%)	28,857 (58.43%)	2,515 (5.09%)
Step 2	7,596 (4.12%)	117,776 (63.99%)	58,686 (31.88%)

either `exact_match` or `partial_match`) query-product annotations. After iterative mining, the number of relevant matches increased to 172,247 out of a total of 233,448 annotations. This represents a 3x increase and shows that Step 2 is critical to the annotation process.

Table 5 lists the differences in distribution of labels we obtained in both steps. We can see that we get a higher proportion of `exact_match` for Step 1 than Step 2 (i.e. 36% vs 4%). Step 2 produces a higher proportion of `partial_match` labels (i.e. 64% vs 58%).

This second step is important, since it can give us more more `exact_match` labels. And also we can view these `partial_match` labels as a possible reflection of the harder-to-score/debatable items on the decision boundary. This increases the difficulty of the dataset to further the discriminative power of the dataset.

8 Conclusions and Future Work

Search engines are critical to the success of e-commerce platforms. Much of the work around the evaluation of these systems tends to be proprietary. We hope that the release of *WANDS* will spur continued research in this domain. In this paper, we described the annotation process we have used in detail, as well as shared evaluation results to showcase the discriminative power of *WANDS*. To recap, our key contributions include: 1) making the dataset available in the public domain, 2) introducing the annotation process and releasing the annotation guidelines we used for reproducibility, and 3) sharing our proposal of cross-referencing as a way to improve dataset completeness while keeping the annotation problem tractable. To the best of our knowledge, *WANDS* is the largest search relevance dataset targeted at e-commerce applications.

Looking ahead, we plan to investigate and compare more approaches for cross-referencing. We also want to confirm our hypothesis that the guidelines we have refined through our annotator training process are sufficient to allow less-trained crowdsourced annotators to produce similarly high-quality datasets.

Acknowledgement

We like to thank Elizabeth Yukman and Alex Wolff for their help with editing this manuscript, John Costello and Ariel Nissan for their legal support, and Natali Vlatko and her team for facilitating the open-sourcing of *WANDS*.

References

1. Home depot product search relevance. https://www.kaggle.com/c/home-depot-product-search-relevance/overview
2. Gov2 dataset. http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm
3. Clueweb09 dataset. http://boston.lti.cs.cmu.edu/Data/ clueweb09/
4. Crowdflower search results relevance. https://www.kaggle.com/c/crowdflower-search-relevance/overview
5. Allan, J., Aslam, J., Carterette, B., Pavlu, V., Kanoulas, E.: Million query track 2008 overview. NIST Special Publication p. 23 (11 2008)
6. Allan, J., Carterette, B., Dachev, B., Aslam, J.A., Pavlu, V., Kanoulas, E.: Million query track 2007 overview. In: Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007 (2007)
7. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 25–32 (2004)
8. Carmel, D., Haramaty, E., Lazerson, A., Lewin-Eytan, L., Maarek, Y.: Why do people buy seemingly irrelevant items in voice product search? on the relation between product relevance and customer satisfaction in ecommerce. In: Proceedings of the 13th International Conference on Web Search and Data Mining. pp. 79–87 (2020)
9. Carterette, B., Pavluy, V., Fang, H., Kanoulas, E.: Million query track 2009 overview. In: TREC (01 2009)
10. Choi, J.I., Kallumadi, S., Mitra, B., Agichtein, E., Javed, F.: Semantic product search for matching structured product catalogs in e-commerce. In: https://arxiv.org/pdf/2008.08180.pdf (2020)
11. Cleverdon, C.: The cranfield tests on index language devices. In: Aslib proceedings. MCB UP Ltd (1967)
12. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement (1960)
13. Deng, A., Shi, X.: Data-driven metric development for online controlled experiments: Seven lessons learned. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 77–86 (2016)
14. Lu, X., Moffat, A., Culpepper, J.S.: The effect of pooling and evaluation depth on ir metrics. Information Retrieval Journal **19**(4), 416–445 (2016)
15. Magnani, A., Liu, F., Xie, M., Banerjee, S.: Neural product retrieval at walmart.com. In: Companion Proceedings of the 2019 World Wide Conference (2019)
16. Nigam, P., Song, Y., Mohan, V., Lkshman, V., Ding, W., Shingavi, A., Teo, C.H., Gu, H., Bing, Y.: Semantic product search. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2019)
17. Pitoura, E., Stefanidis, K., Koutrika, G.: Fairness in rankings and recommendations: an overview. The VLDB Journal (2021)
18. Qin, T., Liu, T.: Introducing LETOR 4.0 datasets. CoRR **abs/1306.2597** (2013), http://arxiv.org/abs/1306.2597
19. Sakai, T., Kando, N.: On information retrieval metrics designed for evaluation with incomplete relevance assessments. Information Retrieval **11**(5), 447–470 (2008)
20. Sanderson, M., Zobel, J.: Information retrieval system evaluation: effort, sensitivity, and reliability. In: Proceedings of the 28th ACM SIGIR Conference on Research and Development in Information Retrieval (2005)

21. Valcarce, D., Bellogín, A., Parapar, J., Castells, P.: On the robustness and discriminative power of information retrieval metrics for top-n recommendation. In: Proceedings of the 12th ACM conference on recommender systems. pp. 260–268 (2018)
22. Voorhees, E.: The sixteenth text retrieval conference (trec 2007) (2007), https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=890068
23. Voorhees, E.M.: The philosophy of information retrieval evaluation. In: Workshop of the Cross-Language Evaluation Forum for European Languages. Springer (2001)
24. Xia, X., Wang, S., Zhang, H., Wang, S., Xu, S., Xiao, Y., Long, B., Yang, W.Y.: Searchgcn: Powering embedding retrieval by graph convolution networks for e-commerce search. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021)
25. Zhang, H., Wang, S., Zhang, K., Tang, Z., Jiang, Y., Xiao, Y., Weipeng, Y., Yang, W.Y.: Towards personalized and semantic retrieval: An end-to-end solution for e-commerce search via embedding learning. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2020)
26. Zhang, H., Wang, T., Meng, X., Hu, Y., Wang, H.: Improving semantic matching via multi-task learning in e-commerce. In: Proceedings of the 42nd International ACM SIGIR on Research and Development in Information and Retrieval, Workshop on eCommerce (2019)
27. Zhang, J., Liu, Y., Ma, S., Tian, Q.: Relevance estimation with multiple information sources on search engine result pages. In: Proceedings of the 2018 ACM on Conference on Information and Knowledge Management (2018)
28. Zhang, J., Liu, Y., Ma, S., Tian, Q.: Que2search: Fast and accurate query and document understanding for search at facebook. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (2021)
29. Zheng, Y., Fan, Z., Liu, Y., Luo, C., Zhang, M., Ma, S.: Sogou-qcl: A new dataset with click relevance label. In: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (2018)