CWRCzech: 100M Query-Document Czech Click Dataset and Its Application to Web Relevance Ranking

Josef Vonášek
Seznam.cz
Prague, Czech Republic
josef.vonasek@firma.seznam.cz

Milan Straka
Charles University, Faculty of Mathematics and Physics
Prague, Czech Republic
straka@ufal.mff.cuni.cz

Rostislav Krč
Seznam.cz
Prague, Czech Republic
rostislav.krc@firma.seznam.cz

Lenka Lasoňová
Seznam.cz
Prague, Czech Republic
lenka.lasonova@firma.seznam.cz

Ekaterina Egorova
Seznam.cz
Prague, Czech Republic
ekaterina.egorova@firma.seznam.cz

Jana Straková
Charles University, Faculty of Mathematics and Physics
Prague, Czech Republic
strakova@ufal.mff.cuni.cz

Jakub Náplava
Seznam.cz
Prague, Czech Republic
jakub.naplava@firma.seznam.cz

ABSTRACT

We present CWRCzech, **C**lick **W**eb **R**anking dataset for **Czech**, a 100M query-document Czech click dataset for relevance ranking with user behavior data collected from search engine logs of Seznam.cz. To the best of our knowledge, CWRCzech is the largest click dataset with raw text published so far. It provides document positions in the search results as well as information about user behavior: 27.6M clicked documents and 10.8M dwell times. In addition, we also publish a manually annotated Czech test for the relevance task, containing nearly 50k query-document pairs, each annotated by at least 2 annotators. Finally, we analyze how the user behavior data improve relevance ranking and show that models trained on data automatically harnessed at sufficient scale can surpass the performance of models trained on human annotated data. CWRCzech is published under an academic non-commercial license and is available to the research community at https://github.com/seznam/CWRCzech.

CCS CONCEPTS

• **Information systems** → **Test collections**; *Web log analysis*; *Relevance assessment*.

KEYWORDS

User Behavior Dataset, Clicks, Dwell times, Relevance Ranking, Web Search, Contrastive Training, Czech

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

ACM Reference Format:
Josef Vonášek, Milan Straka, Rostislav Krč, Lenka Lasoňová, Ekaterina Egorova, Jana Straková, and Jakub Náplava. 2024. CWRCzech: 100M Query-Document Czech Click Dataset and Its Application to Web Relevance Ranking. 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, 11 pages. https://doi.org/10.1145/3626772.3657851

1 INTRODUCTION

In the field of information retrieval (IR), the task of relevance ranking is to determine the degree of relevance of documents or items with respect to a particular query. In order to accommodate lengthier, more naturally phrased queries as opposed to keywords, modern relevance has moved away from rule-based approaches towards pre-trained language models [3, 10, 18, 20, 21, 23, 24]. However, the effectiveness of these models depends heavily on the availability of extensive training data.

Although human relevance annotation provides high-quality training data, it is costly and time-consuming. Harnessing user behavior data collected in production offers a robust, cost-effective option; nevertheless, such query-document click datasets are not routinely published or available for academic non-commercial use at scale, much less so in non-English languages. To date, a few large-scale datasets containing user behavior data in the search domain have been released [9, 27, 29, 38, 39].

In order to contribute to the research area of user behavior in the context of relevance ranking, we publish CWRCzech (**C**lick **W**eb **R**anking dataset for **Czech**), a new dataset of 100M query-document pairs in the Czech language derived from search engine logs of Seznam.cz.[1] It contains not only positive examples but also negative ones (offered but not clicked), which makes it a valuable resource for model training. To our knowledge, the presented dataset is by far the largest click dataset with raw text.

[1] https://search.seznam.cz/

To provide a proper evaluation benchmark, we have also manually annotated and released a representative Czech test set for the relevance task, containing circa 50k query-document pairs, each manually annotated by at least 2 annotators to ensure high quality of the annotations.

To showcase the research potential of the corpus for the relevance ranking research, we analyze and experimentally validate to what degree such automatically collected, inherently noisy user behavior data contribute to training large language models for the relevance ranking task in comparison with human-annotated data. We find that user data generated automatically at sufficient scale challenge the performance of human annotations when evaluated on in-domain relevance ranking. Our contributions are:

- a new, large query-document click dataset CWRCzech for Czech relevance ranking containing 100M query-document pairs of user behavior collected in production,
- manually annotated Czech test set of around 50k query-document pairs, each annotated by at least 2 annotators,
- model analysis and experimental validation of the contribution of automatically harnessed query-document click data for relevance ranking.

CWRCzech is published under a non-commercial license and is available at **https://github.com/seznam/CWRCzech**.

2 RELATED WORK

Datasets. Two primary methods are used to create relevance datasets. The first involves human annotators who manually assess the relevance of each document. The second method directly utilizes user behavior data collected during the use of the company's services.

For instance, Microsoft's MS MARCO dataset [22] includes 8.8 million query-document pairs with human-provided relevance annotations. There are also non-English relevance datasets like the Chinese T2Ranking [32] consisting of 2 million annotated pairs, and the Czech DaReCzech, featuring 1.6 million pairs.

The large-scale English click datasets are often compiled directly from search engine logs. Notable examples are the AOL [25] and the MSN [36] datasets with millions of queries. Microsoft recently released the ORCAS dataset [9] containing 18.8 million query-document pairs. As an example of a domain-specific dataset, one can look at TripClick with 1.3 million pairs [27], acquired from a health web search engine. The most notable non-English datasets provided by other search engine companies include Russian Yandex-WSCD [29] with 35 million search sessions and anonymized queries, or Chinese Sougou-QCL [38] and Baidu-ULTR [39] datasets with 12.2 million pairs and 1.2 billion pairs respectively, with queries and documents anonymized using a proprietary dictionary.

The type of information provided in a click dataset is vital and holds greater importance than size alone when training neural models for retrieval and ranking. A detailed comparison of different click datasets to CWRCzech is discussed in Section 3.1. Notably, click datasets differ in the level of query token anonymization, ranging from dataset replacing all words with randomized IDs [29], through partial word replacement [39], to datasets with original, raw text, as is the case of ORCAS [9] and our dataset. Proper handling of tail, sensitive or harmful queries is important, and we cover this in Section 3.

Relevance Ranking Approaches. Traditional non-neural techniques based on term frequency, such as BM25 [28] or term frequency-inverse document frequency (TF-IDF) representations, were recently widely replaced or extended by neural methods based on the Transformer architecture [30], such as BERT [11]. Pretrained language models have demonstrated their capability for dense retrieval [37] in application domains such as E-commerce [34], recommendation [31], online advertising [33], or web search [4].

Unlabeled data for contrastive learning in web search originate from logged user interactions with the search engine such as clicks or time spent on a particular result (dwell time). These feedback signals can be viewed as an approximation of relevance and used as positive labels. To increase training effectiveness, the larger unlabeled log data can be combined with smaller human-annotated relevance sets for further fine-tuning.

Click data are, for example, frequently coupled with annotated relevance to train rankers [5, 34]. However, clicks also suffer from position bias when items in the top positions receive more exposure and therefore have higher click probability than bottom items [8]. Position bias is commonly considered in models for click-through rate prediction, but it can also improve the quality of training data for relevance models [34].

Another option is to incorporate user's dwell time [16, 35] into the training objective. Short interaction with the clicked result may indicate poor relevance and vice versa. Improved performance when high-quality items are placed in the top positions is observed when clicks are reweighted by the normalized dwell time [31].

3 DATASETS

3.1 CWRCzech Click Dataset

CWRCzech (**C**lick **W**eb **R**anking dataset for **Czech**) is a new Czech click dataset comprising 100 million query-document pairs derived from search engine logs of Seznam.cz. It contains over 2.7 million distinct queries and over 8.4 million documents. The queries come from requests collected over an extended period of time. We only selected the queries in Czech (according to the internal classifier) that were identified by the search engine as an informational intent [6]. Informational intent queries seek to acquire information (e.g., "how to boil an egg") and tend to be more naturally phrased, as opposed to navigational or transactional intent queries. Table 1 provides an example of a search results record for a user query.

When constructing the dataset, our goal was to avoid sensitive information, user identification, and harmful content. To this end, we adhered to the following protocol: All queries classified as porn or obscene were filtered; as well as bot queries. To prevent an accidental leak of numerical information, such as credit card numbers, only queries with alphabetical characters were selected. Sessions were not merged by user ids for privacy reasons. Finally, each query had a minimal occurrence in 5 unique requests within the specified time frame (i.e., the same query was requested by at least 5 users) to ensure anonymization and to prevent potential identification of specific users or their sensitive information.

In order to enhance query variability, the maximal number of unique requests for each query was limited to 15 (i.e., if the same query was requested by more than 15 users, we choose 15 requests at random), yielding 22.1M unique requests over the entire dataset.

Table 1: Visualization of a query from the CWRCzech dataset. The query and the documents have been translated to English for better understanding, and only excerpts of the body text extracts are shown.

REQUESTID 18242939	QUERY automatic parking				
TITLE Automatic parking is not just a privilege of luxury cars - roadblog.cz	URL …	RANK 0	CLICKS 1	DWELLTIME 116	
BTE Arrive at a parking spot, press a button, and let the car park itself. Such a feature is now available in accessible cars as well. You might say, that …					
TITLE Drivers do not use automatic parking, even though it is better than a human – AutoRevue.cz	URL …	RANK 1	CLICKS 0	DWELLTIME 0	
BTE Automatic parking also uses 47 % fewer maneuvers and corrections, and there was not a single instance of contact with another vehicle, unlike …					
TITLE Automatic parking - cars that park themselves | OneTwoGo Car Rental	URL …	RANK 3	CLICKS 0	DWELLTIME 0	
BTE Parallel parking is a struggle for many drivers, especially in big cities. Given that parking space is significantly limited by cars on crowded streets …					
TITLE Automatic Parking - Glossary of Terms - Electric Cars | Alza.cz	URL …	RANK 5	CLICKS 0	DWELLTIME 0	
BTE Most automobile manufacturers provide the feature of automatic parking as an additional equipment option, even for their more affordable …					
TITLE Description and principle of operation of the automatic parking system - AvtoTachki	URL …	RANK 7	CLICKS 1	DWELLTIME N/A	
BTE Parking a car is perhaps the most common maneuver that causes difficulties for drivers, especially inexperienced ones. But it was not so long …					

Table 2: Comparison of CWRCzech to other publicly available click datasets for ranking. The table displays the number of unique queries, documents, total query-document (Q-D) pairs, and the search results sessions. Information about the data contained in each dataset and their languages is provided. *TripClick comes from healthcare domain contrary to other listed web search datasets. §The tokens are represented as identifiers to a private dictionary; therefore, they cannot be used with pre-trained language models. †Dataset contains click-model generated relevance labels. ‡ Dataset contains click sequence, displayed time/count, and others.

Dataset	Q-D pairs	Queries	Docs	Language	Query text	Doc title	Doc body	Clicks	Dwell time	Rank	Additional Information
ORCAS [9]	18.8M	10.4M	1.4M	English	raw	raw	raw	✓	-	-	-
TripClick [27]	5.3M	1.6M	2.3M	English*	raw	raw	raw	✓	-	✓	-
TianGong-ULTR [1, 2]		3.4K	333.8K	Chinese	raw	raw	raw	✓	-	✓	-
Sougou-QCL [38]	12.2M	0.5M	9.0M	Chinese	raw	raw	raw	-	-	-	✓†
Baidu-ULTR [39]		383.4M	1.3B	Chinese	private§	private§	private§	✓	✓	✓	✓‡
Yandex-WSCD [29]	667.2M	21.1M	70.3M	Russian	private§	-	-	✓	-	✓	-
CWRCzech	100.0M	2.7M	8.4M	Czech	raw	raw	≤230 chrs	✓	✓	✓	-

For inclusion in the dataset, each request was associated with documents extending up to the last click or up to the fifth position, whichever was greater. To concentrate on more complex inquiries, the dataset was curated to include only queries comprising a minimum of 10 characters.

The dataset contains the following columns:

- *requestId*: Id of the particular request with a single query.
- *query*: User query with corrected typos and added diacritical marks.
- *url*: Document URL.
- *title*: Words from the document classified by the search engine as a title.
- *bte*: Body text extract, i.e., document body snippet processed by the internal search engine model and trimmed to 230 characters (snippet size complying with fair use). It is empty for the webpages that block search engines or prohibit usage of their contents for GPT training.
- *rank*: Position of the document in the search results page. Indices may be absent in cases where a document was no longer indexed at the time of dataset creation.
- *clicks*: The number of clicks on a given document in given search results.
- *dwellTime*: Time in seconds spent in the clicked document page before the user returned to the search results page. This information is not always available, typically for the last click in the search results.

Table 2 presents statistics summarizing the new CWRCzech dataset in comparison with other click datasets. Among datasets with readable text, i.e., without identifiers into a private dictionary, CWRCzech is the largest one, thus it is significantly larger than ORCAS [9]. The total number of documents is higher, however, it has a lower count of unique queries. The median query length in CWRCzech is 3 words and the average length is 3.48 words. Contrary to ORCAS, long queries and one-word queries are less

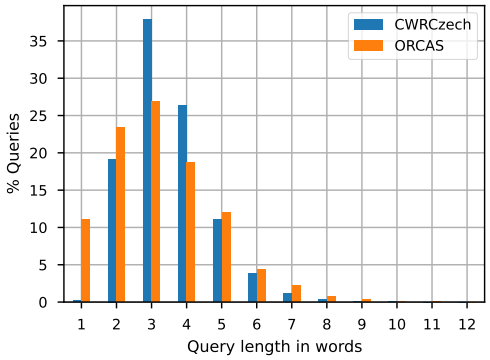


Figure 1: Comparison of query length in words (separated by a blank space) between CWRCzech and ORCAS [9].

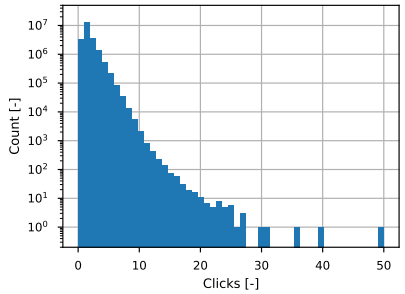
common in CWRCzech. The distribution is illustrated in Figure 1. Moreover, unlike ORCAS, CWRCzech contains whole search results with both clicked and unclicked documents as well as information about document rank and user dwell time. This provides a detailed insight into user interaction with search results that can be utilized for more precise relevance estimates and efficient ranker training.

Figure 2 illustrates some of the CWRCzech statistics: the distribution of the number of clicks per query-document pair, the distribution of dwell times where they are available, the number of documents per query, and the correlation between the rank of the document and the number of clicks it receives. Out of all individual query-document pairs, 27.6M are clicked; after aggregation, 60% of the query-document pairs received no clicks, 24% were clicked exactly once, and 16% of the documents received more than one click (see Figure 2.a). Both clicks and non-clicks provide valuable information about user interaction with the search results. Clicks are commonly treated as a strong relevance signal, but non-clicks can be used, for example, for the construction of soft negative pairs (see Section 4.5). User post-click behavior is equally significant. An example of such behavior is dwell time, which is provided in CWRCzech explicitly. Note that it is available for 10.8% pairs with the mean value of 132.5 seconds and the median of 58 seconds (see Figure 2.b). The rank of the document on a page provides an insight into a potential position bias. Number of documents per query peaks at 10 (Figure 2.c) which is the size of the first results page. The probability of a click generally decreases with increasing rank and more than half of the clicks occur in the top three results (see Figure 2.d). Results paging causes visible steps in the graph for ranks divisible by the search results length since users are more likely to click on top documents on each page.

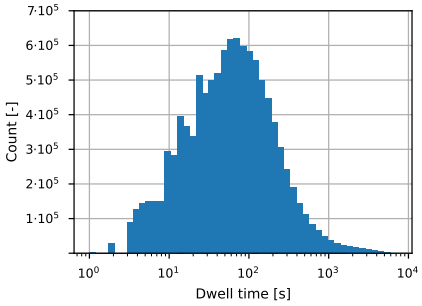
As all European-based companies are required to comply with the General Data Protection Regulation (GDPR), most relevantly Act. 17 Right to Erasure ('Right to be Forgotten'), the corpus is available under a non-commercial license upon request to keep record of the corpus users for broadcasting the potential erasure requirements.

3.2 CWRCzech Annotated Test Set

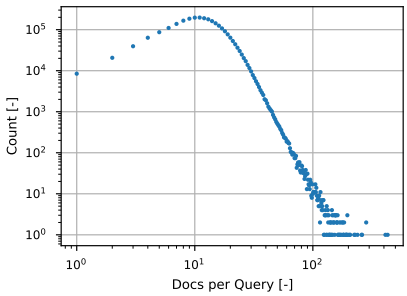
Along with the CWRCzech click dataset, we also publish a manually annotated Czech test set for model evaluation.



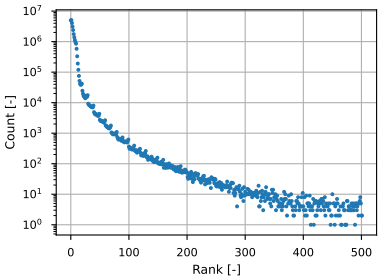
(a) Number of clicks per requestId.



(b) Distribution of dwell times for clicked documents.



(c) Number of documents per query.



(d) Distribution of ranks for clicked documents.

Figure 2: Statistics of clicks, dwell times, and ranks within the CWRCzech dataset.

We retrieved queries from the 2023 search logs, and performed random selection, deduplication, filtering of only informational intent as in Section 3.1. Each query was manually scrutinized for safety, sensibility, and anonymity. This process resulted in 995 queries that were designated as the test set. Each query is paired

Table 3: Annotated test sets data summary. The table displays the total number query-document (Q-D) pairs, the number of unique queries, the average number of documents per query (Avg. D/Q), and the number of annotators per query-document pair (Num. Annots).

Test Set Name	Q-D pairs	Unique Queries	Avg. D/Q	Num. Annots
DaReCzech (test)	64,466	2,323	27.75	1
- *non-informational intents*	54,899	1,609	34.12	1
DaReCzech (dev)	41,220	793	51.98	1
- *informational intent*	4,828	99	48.77	1
CWRCzech	49,945	995	50.20	2-3

with documents of varying relevance, ranging from highly relevant documents in the search results to those with little to no relevance, often found in the early stages of retrieval. The texts of the documents were trimmed to 230 characters to comply with fair use. In total, the test set comprises 49,945 rows of query-document pairs. On average, each query is linked to 50.20 documents, with a minimum of 31 and a maximum of 89, and 19.27% of these documents are deemed relevant ($label > 0.5$, see Label Design in Section 4.4). There is no overlap between the annotated test set and the CWRCzech click set.

Each query-document pair underwent annotation by at least 2 annotators,[2] who had the option to assign the pair one of the following four values indicating the usefulness of the document with respect to the query: "useful" (labeled as 1), "slightly useful" (0.66), "almost useless" (0.33), and "useless" (0). In case when two annotations considerably differed, i.e., one was "(slightly) useful" and the other "(slightly) useless", a third annotator was asked to provide another annotation. This happened in 10% of the cases. The ultimate label for each pair was determined by calculating the median of the two or three assigned values.

3.3 DaReCzech Dataset

In addition to the aforementioned test set, our experiments also make use of a previously released Czech dataset DaReCzech [18]. Like the CWRCzech test set, the DaReCzech test set is also human annotated. However, its annotation reliability is lower than that of CWRCzech as the labels are based on a single annotation for each query-document pair. Another key distinction between the two datasets is that DaReCzech encompasses a full range of user intents, not just informational ones. As shown in Table 3, there is a significant domain shift in intents between CWRCzech and DaReCzech test set, as circa 70% of the DaReCzech test is non-informational intent. Hence, in this paper, we use the DaReCzech test set for out-of-domain robustness testing (Section 6). DaReCzech development queries with informational intent were allocated to the development set that is used as the stopping criterion during training. Both the DaReCzech development and test set as well as other parameters are shown in Table 3.

[2]Our annotators were in-house expert employees, native speakers of Czech, and predominantly women. They were compensated based on the number of annotations they made and their pay was above the mean salary valid in 2023 in the relevant country.

4 METHODOLOGY

To demonstrate the potential contribution of user behavior data as a complement or replacement of human annotations, we finetune three encoder-only pretrained language models for a relevance ranking task in cross-encoder and bi-encoder settings [26], because both have their specific uses in web search. The inputs consist of a query-document pair, with the document represented by its url, title, and text. We use the aggregated user behavior (clicks and dwell times) along with the document positions from CWRCzech in order to construct pseudo labels as approximates of human annotations.

4.1 Architectures

We conduct training for each model using both cross-encoder and bi-encoder [26] configurations. Cross-encoders generally yield superior outcomes in web ranking tasks; however, their slower processing speed renders them impractical for production environments, where rapid ranking of thousands or millions of query-document pairs is essential. Bi-encoders circumvent this performance bottleneck by computing the embeddings independently for queries and documents. Nonetheless, cross-encoders retain their significance in the training phase, acting as effective teachers during the training process of bi-encoders.

4.1.1 Cross-encoder. The term cross-encoder [26] describes an architecture that follows the original approach for sequence classification, e.g., BERT [11]. As illustrated in Figure 4, the input to this model is a single sequence – a query and a document separated by the special `[SEP]` token. To predict the relevance of a query to a document, we add an additional linear layer (FFN) on top of the classification token (`[CLS]`) embedding with a sigmoid activation ($label_{pred}$) to project a score between 0 and 1.

4.1.2 Bi-encoder. The bi-encoder siamese architecture [26] (illustrated in Figure 3, together with the loss computation described later in this section) utilizes an identical pre-trained model to compute the embeddings of the query and the document separately. The embeddings are then processed by the interaction module ("Head" in Figure 3) introduced by Kocián et al. [18], which first computes the element-wise maximum of the embeddings and passes it through a 2-layer feed-forward network with a residual connection, concatenates the output with the Euclidean distances and cosine similarity of the two embeddings, and passes the result through a final linear transformation with a sigmoid activation.

4.2 Pretrained Models

We finetune three encoder-only pretrained Czech models: Small-E-Czech [18], RetroMAE-Small [4], and Fernet-C5 [19]. The model parameters are presented in Table 4.

Of these models, only RetroMAE-Small has been pretrained to produce high-quality sentence embeddings [4], making it particularly well-suited for the bi-encoder architecture. Nonetheless, we demonstrate that the other models without a sentence-embedding pretraining objective can attain competitive performance in the bi-encoder settings.

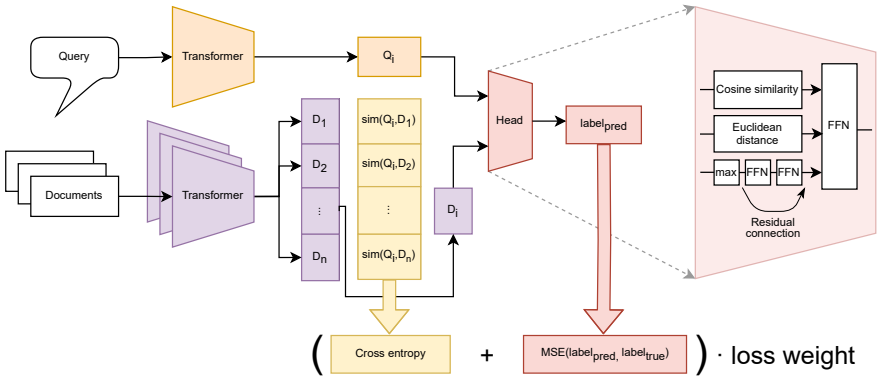


Figure 3: A visualization of the proposed architecture of a bi-encoder relevance model.

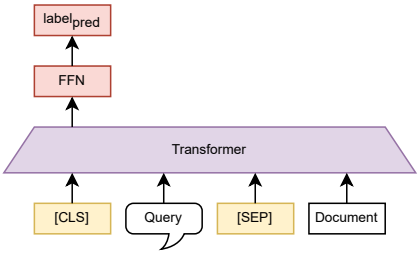


Figure 4: An illustration of a cross-encoder relevance model.

Table 4: Model sizes and corresponding parameters. For each model, the total number of parameters is given, and in the brackets is the number of parameters in the embedding layer and the rest of the parameters.

Model	Size [type]	Number of params.
Small-E-Czech [18]	Small	13M (4M+ 9M)
RetroMAE-Small [4]	Small	24M (9M+15M)
FERNET-C5 [19]	Base	162M (85M+77M)

4.3 Training Details

We train the models utilizing the Adam optimizer [17] using the default hyperparameters and no weight decay. We use accumulated batch size of 500, and learning rates 5e-5 and 2e-5 for small and base models, respectively. For CWRCzech, we train the models for 4 epochs. For DaReCzech, we increase the number of epochs to 10. With this setup and 4 A100 GPUs, we are able to train a single small cross-encoder model in approximately 1 day, and a single base cross-encoder model in approximately 4 days. The training time is doubled for bi-encoders. Finally, we select the epoch with the highest NDCG@10 performance on DaReCzech dev set queries restricted to informational intent (Section 3.3).

4.4 Label Design

CWRCzech contains various user behavior information, however, the way of constructing a label as a target model variable for training from this information is not readily apparent. In this Section, we show how clicks, dwell times, and document positions can be used as a relevance label proxy. Ultimately, we propose the resulting most effective formula to combine these three attributes into a single pseudo relevance label, motivated by the correlation analysis carried out in Table 5.

Table 5: Spearman correlation between various aspects of user behavior and manual annotations on the DaReCzech train set. Formulas in Sections 4.4.5, 4.4.6 are used for rank and click & dwell time & rank, respectively.

Behavior type	Correlation
Rank	0.0755
Dwell Time (NaN→0)	0.0884
Clicks	0.1335
Dwell Time (NaN→mean)	0.1419
Dwell Time (NaN→mean) × (Clicks + Rank)	0.1463

4.4.1 Aggregation. Many query-document pairs occur multiple times in CWRCzech. Therefore, when constructing a label for a single query-document pair, we need to aggregate all information available for this pair. To that end, we sum all clicks, dwell times, and ranks for every unique query-document pair, and consider only these sums in the rest of the section. The use of sum over mean or median might seem counter-intuitive, but based on our experiments, sum outperformed other aggregation methods.

4.4.2 Correlations. We start by computing the Spearman rank correlation of aggregated clicks, dwell times, click ranks, and their combination with the reference annotations. The reference annotations are obtained by searching all query-document pairs from CWRCzech in the DaReCzech train set.

The results are presented in Table 5. We see that all considered values show small positive correlation with the reference annotations. The lower correlation of the raw dwell times with reference annotations is partly caused by missing dwell times for 89.2% of the pairs, because zero is used for missing values. When mean dwell times are substituted for the missing values, the correlation surpasses clicks. Finally, the combination *click & dwell time & rank* based on the formula in Section 4.4.6 shows the best correlation with the annotations.

4.4.3 Clicks. When considering clicks, we distinguish between a *last* click in a request and *nonlast* clicks, because the *last* click may indicate either that the user found the required information or abandoned the search. We therefore define weighted clicks as

$$wclicks(q,d) \leftarrow \alpha \cdot nonlast\_clicks(q,d) + \beta \cdot last\_clicks(q,d),$$

where α and β are the weights of nonlast and last clicks, respectively.

To construct labels from weighted clicks, we need to map them to a label in the $[0,1]$ range. We considered monotone mappings, assuming more clicks signify higher relevance, and a log transformation with a suitable scale delivered the best performance on DaReCzech information-intent development set. We therefore assign the following label to given weighted clicks:

$$l(q,d) \leftarrow \Big| s \cdot \log\big(1 + wclicks(q,d)\big) \Big|_0^1,$$

where $l(q,d)$ is the generated label for query q and document d, $|x|_0^1 = \max(0, \min(1,x))$ clips the input value to the interval $[0,1]$, and s is a scale factor we describe in Section 4.4.6.

4.4.4 Dwell time. Dwell time is known only for 39% of clicks in our dataset, but where available, it provides additional valuable information about user engagement. Analogously to clicks, we obtain the label by transforming the aggregated total dwell time by the log function:

$$l(q,d) \leftarrow \Big| s \cdot \log\big(1 + dwelltimes(q,d)\big) \Big|_0^1.$$

4.4.5 Rank. The effect of the rank of a document on its relevance might be equivocal. On one hand, the search engine aims to generate the most relevant documents on top, implying lower rank should indicate higher relevance; on the other hand, mitigating position bias results in increasing relevance for documents with larger rank [8]. Both the correlations and the trained model performance show that the former effect is stronger (lower rank indicating higher relevance), with the following label calculation delivering best results out of the alternatives we considered:

$$l(q,d) \leftarrow \frac{views(q,d)}{ranks(q,d) + C}.$$

The resulting label is a reciprocal of mean rank, with an additional constant C (experimentally chosen as $C = 100$) to boost more frequent documents.

4.4.6 ClickDwellRank. We also consider combining clicks, dwell times, and rank into a single label. The empirically most successful formula found in our preliminary experiments is the following:

$$l(q,d) \leftarrow \left| s \cdot \log\left(1 + \left(wclicks(q,d) + \frac{views(q,d)}{ranks(q,d)+C}\right) \cdot \Big| dwelltimes(q,d) \Big|_1^\infty \right) \right|_0^1.$$

The ranks in this formula work mostly as a tie breaker, in case when two documents receive the same amount of clicks and dwell time. The document with higher position (lower rank) then obtains a higher label.

Finally, we set the scale factor s so that virtually all labels are less than 1 and do not need to be clipped, choosing $s = 1/20$. For simplicity, we employ the same scale factor for all configurations.

4.4.7 Loss Weights. Given that our labels are aggregated across all query-document pairs, we lose the information about prevalence of each such pair. We can restore the information by using loss weights – for each query-document pair, we multiply its loss by its number of occurrences in the dataset:

$$loss(q,d) \leftarrow loss(q,d) \cdot \log\big(2 + views(q,d)\big).$$

Apart from the number of occurrences, we could also try to mitigate the natural imbalance between clicks and non-clicks, given that clicks account for only 27.6% of query-document pairs in our dataset. Therefore, we also examine an alternative formula accentuating clicked results:

$$loss(q,d) \leftarrow loss(q,d) \cdot \log\big(2 + clicks(q,d)\big).$$

The constant 2 is needed in order for the loss weight to be strictly positive even when there are no clicks.

4.5 Soft Negative Pairs

Because the documents in CWRCzech are retrieved by a search engine, they are expected to be highly relevant to a given query. However, for successful training, a model might also benefit from clearly non-relevant documents (with relevance label 0). We call such query-document pairs the *soft* negative examples.

A straightforward approach to generate soft negative pairs is to randomly sample a constant number of documents for every query. This approach is naturally applicable in both the cross-encoder and the bi-encoder setting.

4.5.1 Contrastive Training. In the bi-encoder setting, we can obtain the soft negatives for a given query more efficiently by considering all documents relevant to other queries in the batch – the so-called *in-batch* negative examples [14]. Specifically, we follow the standard contrastive framework [7] and use the cross-entropy objective with in-batch negatives as a contrastive loss:

$$\text{constrastive loss}(q,d) \leftarrow -\log \frac{e^{\mathrm{sim}(q_i,d_i)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(q_i,d_j)/\tau}}.$$

Here, N is the number of documents in the batch, τ is a temperature parameter that controls the separation of positive and negative examples (our method uses learnable τ with an initial value of 0.07), and $\mathrm{sim}(q,d)$ is the similarity measure between the representations of a query q and a document d. We specifically use the cosine similarity, hence

$$\mathrm{sim}(q,d) = \frac{q^T d}{\|q\| \cdot \|d\|}.$$

5 RESULTS

Our best model uses the label formula presented in Section 4.4.6 along with generated soft negative query-document pairs (Section 4.5) and click-based loss weights (Section 4.4.7). Furthermore, our bi-encoders also use the contrastive training objective based on the cross-entropy loss with initial temperature $\tau = 0.07$ as described in Section 4.5.1. We report model performance in three settings:

- DaReCzech: training solely on human-labeled dataset,
- CWRCzech: training solely on our click dataset,
- CWRCzech + DaReCzech: pretraining on clicks and finetuning on human annotations.

Table 6: Baseline NDCG@10 [%] on the CWRCzech test set.

Dataset	Random baseline	Oracle baseline
CWRCzech test	22.50	98.69

Table 7: Evaluation of the Czech models finetuned on the human annotated DaReCzech train set, the CWRCzech train set, or both. We show an average and a standard deviation of three runs, and $A < B$ indicates the difference between models trained on different train sets (comparing columns) is statistically significant (p less than 0.001), using the Monte Carlo permutation test with 1M samples and probability of error at most 10^{-6} [12, 13].

Test Set:	CWRCzech NDCG@10 [%]		
Train Set:	DaReCzech	CWRCzech	CWRCzech DaReCzech
Bi-encoders:			
Small-E-Czech	62.09 ± 0.2	< 68.01 ± 0.1	68.43 ± 0.1
RetroMAE-Small	64.12 ± 0.2	< 68.73 ± 0.1	68.75 ± 0.2
FERNET-C5	63.09 ± 0.7	< 69.88 ± 0.1	69.89 ± 1.1
Cross-encoders:			
Small-E-Czech	66.41 ± 0.5	< 68.12 ± 0.5	< 69.86 ± 0.8
RetroMAE-Small	65.20 ± 0.4	< 67.48 ± 0.7	< 68.70 ± 0.4
FERNET-C5	68.06 ± 1.1	< 70.05 ± 0.6	< 72.35 ± 0.3

We use the standard ranking metric NDCG@10 [15] and treat only labels with numeric value above 0.5 as relevant. For every configuration, we train a model three times with different random initialization, and report the average and standard deviation.

The main results are presented in Table 7 and for reference, the random and oracle baselines are also supplied in Table 6. For cross-encoders, our method of click training achieves +2 percent points in NDCG@10 over our human-labeled DaReCzech baseline. In case of bi-encoders, this difference becomes even more prominent, showing +4.5 percent point NDCG@10 improvement. These differences are statistically significant with p-value less than 0.001, using the Monte Carlo permutation test with 1M samples and probability of error at most 10^{-6} [12, 13]. When further finetuning the CWRCzech-trained models on DaReCzech, we see further statistically significant improvements in the cross-encoder setting, but nearly no change in the bi-encoder setting. We note that our strong baselines achieve performance comparable to state of the art, as evidenced by Table 10.

An interesting question is at what scale the automatically collected data (CWRCzech) can match the results of human-annotated data (DaReCzech). We show the relation between the amount of click data and performance of our Small-E-Czech cross-encoder model in Figure 5. For each dataset size, we sub-sample a random set of queries together with their relevant documents. The graph clearly shows that in order to match the performance on 1M manual annotations (orange line, DaReCzech), we need around 20M user behavior data (blue line, CWRCzech), so that both systems trained on their respective datasets reach the same NDCG@10 (circa 66.5; note that the x axis is logarithmic). Furthermore, there appears to

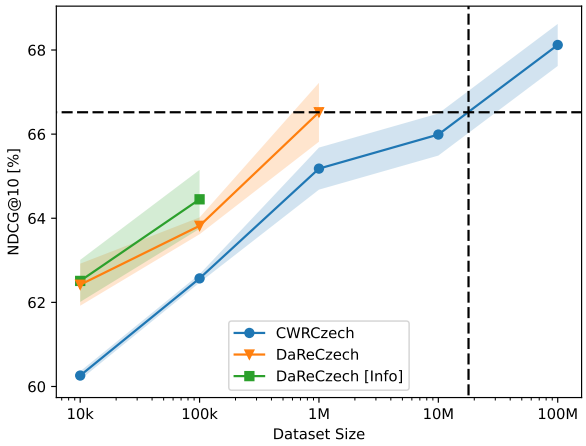


Figure 5: Relationship between user behavior dataset size and Small-E-Czech cross-encoder performance on the CWRCzech test set. The x-axis represents number of query-document pairs in the dataset used for training.

be a roughly linear trend, where each order of magnitude increase in training data size results in approximately a 2 percentage points gain in NDCG@10.

6 ABLATION STUDIES

In this section, we study each component of our final method in isolation.

Labels. In Section 4.4, we detailed various methods for label generation, utilizing user clicks, dwell times, document rank, and a combination of these factors. Additionally, for click-based labeling, it is necessary to choose coefficients (α, β) for nonlast and last clicks, respectively. For label ablation study, we considered the combinations (1, 1), (0.5, 1), and (1, 0.5), thereby assigning varying levels of importance to the last click relative to the rest. We then train a baseline cross-encoder Small-E-Czech model using each designed label to analyze their respective impacts.

The results are presented in Table 8.a, ordered by model performance. Our findings indicate that all click- and dwell time-based labels surpass the baseline model which is trained exclusively on position data (row "Rank"). Notably, the coefficient assigned to the last click plays a significant role, with the best performance surprisingly coming from a model that attenuates its importance (row "Clicks(1, 0.5)"). Ultimately, the combined label approach (row "ClickDwellRank") yields the highest improvement of +1.9 percent points over the position-based label.

Soft Negative Pairs. We demonstrate how extension of the training dataset with automatically generated soft negatives enhances the cross-encoder Small-E-Czech performance in Table 8.b. The effect is substantial despite the simplicity of the method, yielding a +2 percent point increase in NDCG@10.

Contrastive Training. The effect of employing the additional contrastive loss (Section 4.5.1) in the bi-encoder setting is quantified in Table 9.a. Training with additional contrastive loss yields +6 percent points NDCG@10 increase compared to a non-contrastive baseline.

Table 8: Ablation experiments of the Small-E-Czech cross-encoder performance measured on the CWRCzech test set. We show an average and a standard deviation of three runs. (a) The comparison of label design methods (Section 4.4). (b) The influence of soft negative pairs (Section 4.5).

Test Set:	CWRCzech NDCG@10 [%]	
Train Set:	CWRCzech	CWRCzech DaReCzech
(a) QueryDoc Label:		
Rank	65.34 ± 0.3	66.30 ± 0.6
Dwell Time (NaN→20)	65.58 ± 0.9	68.89 ± 0.4
Clicks(0.5, 1)	66.18 ± 0.2	67.12 ± 0.4
Clicks(1, 1)	66.30 ± 0.2	67.46 ± 0.4
Clicks(1, 0.5)	66.51 ± 0.5	67.93 ± 0.8
ClickDwellRank	67.22 ± 0.5	69.30 ± 0.2
(b) Soft Negatives:		
Not included	66.30 ± 0.2	67.46 ± 0.4
Included	68.35 ± 0.2	69.39 ± 0.3

Table 9: Small-E-Czech bi-encoder performance trained with (a) several contrastive objectives (Section 4.5.1) and (b) different loss weights (Section 4.4.7).

Test Set:	CWRCzech NDCG@10 [%]	
Train Set:	CWRCzech	CWRCzech DaReCzech
(a) Contrastive Training:		
None	60.22 ± 0.8	63.72 ± 0.6
Cross-Entropy w/o Head	66.17 ± 0.4	66.92 ± 0.5
Cross-Entropy	66.62 ± 0.5	66.87 ± 0.3
Cross-Entropy Soft Negatives	67.73 ± 0.6	67.70 ± 1.0
(b) Loss Weights:		
None	67.73 ± 0.6	67.70 ± 1.0
Views	68.21 ± 0.2	68.45 ± 0.4
Clicks	68.05 ± 0.2	68.59 ± 0.5

Interestingly, extending the dataset with generated soft negatives as in the cross-encoder settings improves the results further by additional +1 percent points.

Loss Weights. Since we use aggregated behavior labels, the information about original query-document frequency is lost during training. To restore it, we weight the loss function by the number of views a query-document received as described in Section 4.4.7. Furthermore, we also consider number of clicks as a loss weight to mitigate the inbalance between clicked and nonclicked documents.

The results are compared in Table 9.b and show that both approaches improve the model performance. Particularly, using views as weights yields the best improvement of +0.5 percent points.

DaReCzech Evaluation. Since our behavioral data were exclusively collected for informational intent, it is reasonable to verify their robustness and applicability in the out-of-domain setting. For this purpose, we utilize the DaReCzech test set (Section 3.3) evaluated using the P@10 metric employed in prior work [4, 18]. Table 10 indicates that our models reach the performance comparable to prior work when finetuned on DaReCzech, lending credibility to our primary results. Moreover, it also demonstrates that pretraining on user behavior data enhances performance, even when assessed on data outside the original domain.

Table 10: Out-of-domain performance comparison of models on DaReCzech using the P@10 metric, including prior work in italic.

Test Set:	DaReCzech P@10 [%]	
Train Set:	DaReCzech	CWRCzech DaReCzech
Bi-encoders:		
Small-E-Czech [18]	*45.26 ± 0.2*	
RetroMAE-Small [4]	*45.29 ± 0.3*	
FERNET-C5 [4]	*45.87 ± 0.3*	
Small-E-Czech	45.40 ± 0.0 <	46.19 ± 0.2
RetroMAE-Small	45.69 ± 0.1 <	46.38 ± 0.1
FERNET-C5	45.37 ± 0.2 <	46.45 ± 0.4
Cross-encoders:		
Small-E-Czech [18]	*46.30 ± 0.2*	
Small-E-Czech	46.26 ± 0.1	46.43 ± 0.3
RetroMAE-Small	46.28 ± 0.0	46.53 ± 0.2
FERNET-C5	46.95 ± 0.2 <	47.40 ± 0.1

7 CONCLUSION

We introduced a new click dataset for web relevance ranking in Czech, called CWRCzech. It features 100M query-document pairs, of which 27.6M are recorded clicks and 10.8M have dwell times, making it a unique resource of this magnitude. Along with the automatically harnessed user behavior data, we also publish a manually annotated test set with nearly 50k query-document pairs. The dataset is available for academic non-commercial use upon request and is subject to license agreement to ensure compliance with GDPR.

We also carried out extensive experiments comparing various ways of leveraging user behavior data from the corpus for relevance ranking. Our best model uses a combination of clicks, dwell times, and document rank as a target output variable. It also utilizes generated soft negative query-document pairs for contrastive training, and employs click-based loss weights. This model trained on user behavior data from CWRCzech achieves 2.5 percent point improvement for cross-encoder training and 4 percent point for bi-encoder training compared to the baseline trained on human annotated data.

Our analysis of the usefulness of the automatically generated data concludes that for Czech relevance ranking, performance on 1M manually annotated data can be matched by roughly 20M of user behavior data and surpassed with higher quantities.

ACKNOWLEDGMENTS

This work was partially supported by the Grant Agency of the Czech Republic under the EXPRO program (project No. GX20-16819X).

REFERENCES

[1] Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W. Bruce Croft. 2018. Unbiased Learning to Rank with Unbiased Propensity Estimation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. ACM, New York, NY, USA, 385–394. https://doi.org/10.1145/3209978.3209986

[2] Qingyao Ai, Jiaxin Mao, Yiqun Liu, and W. Bruce Croft. 2018. Unbiased Learning to Rank: Theory and Practice. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. ACM, New York, NY, USA, 2305–2306. https://doi.org/10.1145/3269206.3274274

[3] Zeynep Akkalyoncu Yilmaz, Shengjin Wang, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Applying BERT to Document Retrieval with Birch. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, Sebastian Padó and Ruihong Huang (Eds.). Association for Computational Linguistics, Hong Kong, China, 19–24. https://doi.org/10.18653/v1/D19-3004

[4] Jiří Bednář, Jakub Náplava, Petra Barančíková, and Ondřej Lisickỳ. 2024. Some Like It Small: Czech Semantic Embedding Models for Industry Applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22734–22742.

[5] Lin Bo, Liang Pang, Gang Wang, Jun Xu, XiuQiang He, and Ji-Rong Wen. 2021. Modeling Relevance Ranking under the Pre-training and Fine-tuning Paradigm. *CoRR* abs/2108.05652 (2021). arXiv:2108.05652 https://arxiv.org/abs/2108.05652

[6] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (sep 2002), 3–10. https://doi.org/10.1145/792550.792552

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[8] Aleksandr Chuklin, Ilya Markov, and Maarten Rijke. 2015. *Click Models for Web Search*. Springer International Publishing, Cham, Switzerland. https://link.springer.com/book/10.1007/978-3-031-02294-4#bibliographic-information

[9] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 18 Million Clicked Query-Document Pairs for Analyzing Search. In *CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 2983–2989. https://doi.org/10.1145/3340531.3412779

[10] Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 985–988. https://doi.org/10.1145/3331184.3331303

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] Michael P Fay and Dean A Follmann. 2002. Designing Monte Carlo Implementations of Permutation or Bootstrap Hypothesis Tests. *The American Statistician* 56, 1 (2002), 63–70. https://doi.org/10.1198/000313002753631385 arXiv:https://doi.org/10.1198/000313002753631385

[13] Axel Gandy. 2009. Sequential Implementation of Monte Carlo Tests With Uniformly Bounded Resampling Risk. *J. Amer. Statist. Assoc.* 104, 488 (2009), 1504–1511. https://doi.org/10.1198/jasa.2009.tm08368 arXiv:https://doi.org/10.1198/jasa.2009.tm08368

[14] Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning Dense Representations for Entity Retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Mohit Bansal and Aline Villavicencio (Eds.). Association for Computational Linguistics, Hong Kong, China, 528–537. https://doi.org/10.18653/v1/K19-1049

[15] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (oct 2002), 422–446. https://doi.org/10.1145/582415.582418

[16] Youngho Kim, Ahmed Hassan, Ryen W. White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *WSDM '14: Proceedings of the 7th ACM international conference on Web search and data mining*. Association for Computing Machinery, New York, NY, USA, 193–202. https://doi.org/10.1145/2556195.2556220

[17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[18] Matěj Kocián, Jakub Náplava, Daniel Štancl, and Vladimír Kadlec. 2022. Siamese BERT-Based Model for Web Search Relevance Ranking Evaluated on a New Czech Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (Jun. 2022), 12369–12377. https://doi.org/10.1609/aaai.v36i11.21502

[19] Jan Lehečka and Jan Švec. 2021. *Comparison of Czech Transformers on Text Classification Tasks*. Springer International Publishing, 27–37. https://doi.org/10.1007/978-3-030-89579-2_3

[20] Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2023. PARADE: Passage Representation Aggregation for Document Reranking. *ACM Trans. Inf. Syst.* 42, 2, Article 36 (sep 2023), 26 pages. https://doi.org/10.1145/3600088

[21] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized Embeddings for Document Ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1101–1104. https://doi.org/10.1145/3331184.3331317

[22] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. (November 2016). https://www.microsoft.com/en-us/research/publication/ms-marco-human-generated-machine-reading-comprehension-dataset/

[23] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[24] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy J. Lin. 2019. Multi-Stage Document Ranking with BERT. *ArXiv* abs/1910.14424 (2019). https://api.semanticscholar.org/CorpusID:207758365

[25] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st international conference on Scalable information systems*. 1–es.

[26] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[27] Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. 2021. TripClick: The Log Files of a Large Health Web Search Engine. In *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 2507–2513. https://doi.org/10.1145/3404835.3463242

[28] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994 (NIST Special Publication, Vol. 500-225)*, Donna K. Harman (Ed.). National Institute of Standards and Technology (NIST), 109–126. http://trec.nist.gov/pubs/trec3/papers/city.ps.gz

[29] Pavel Serdyukov, Georges Dupret, and Nick Craswell. 2014. Log-based personalization: the 4th web search click data (WSCD) workshop. In *WSDM '14: Proceedings of the 7th ACM international conference on Web search and data mining*. Association for Computing Machinery, New York, NY, USA, 685–686. https://doi.org/10.1145/2556195.2556207

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[31] Ruobing Xie, Lin Ma, Shaoliang Zhang, Feng Xia, and Leyu Lin. 2023. Reweighting Clicks with Dwell Time in Recommendation. In *WWW '23 Companion: Companion Proceedings of the ACM Web Conference 2023*. Association for Computing Machinery, New York, NY, USA, 341–345. https://doi.org/10.1145/3543873.3584624

[32] Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, and Jin Ma. 2023. T2Ranking: A large-scale Chinese Benchmark for Passage Ranking. arXiv:2304.03679 [cs.IR]

[33] Junhan Yang, Zheng Liu, Bowen Jin, Jianxun Lian, Defu Lian, Akshay Soni, Eun Yong Kang, Yajun Wang, Guangzhong Sun, and Xing Xie. 2021. Hybrid Encoder: Towards Efficient and Precise Native AdsRecommendation via Hybrid Transformer Encoding Networks. *CoRR* abs/2104.10925 (2021). arXiv:2104.10925 https://arxiv.org/abs/2104.10925

[34] Shaowei Yao, Jiwei Tan, Xi Chen, Keping Yang, Rong Xiao, Hongbo Deng, and Xiaojun Wan. 2021. Learning a Product Relevance Model from Click-Through Data in E-Commerce. In *WWW '21: Proceedings of the Web Conference 2021*. Association for Computing Machinery, New York, NY, USA, 2890–2899. https://doi.org/10.1145/3442381.3450129

[35] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond clicks: dwell time for personalization. In *RecSys '14: Proceedings of the 8th ACM Conference on Recommender systems*. Association for Computing Machinery, New York, NY, USA, 113–120. https://doi.org/10.1145/2645710.2645724

[36] Yuye Zhang and Alistair Moffat. 2006. Some Observations on User Search Behaviour. *Aust. J. Intell. Inf. Process. Syst.* 9, 2 (2006), 1–8.

[37] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2022. Dense Text Retrieval based on Pretrained Language Models: A Survey. arXiv:2211.14876 [cs.IR]

[38] Yukun Zheng, Zhen Fan, Yiqun Liu, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Sogou-QCL: A New Dataset with Click Relevance Label. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 1117–1120. https://doi.org/10.1145/3209978.3210092

[39] Lixin Zou, Haitao Mao, Xiaokai Chu, Jiliang Tang, Wenwen Ye, Shuaiqiang Wang, and Dawei Yin. 2022. A Large Scale Search Dataset for Unbiased Learning to Rank. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 1127–1139. https://proceedings.neurips.cc/paper_files/paper/2022/file/07f560092a0edceabf55af32a40eaee3-Paper-Datasets_and_Benchmarks.pdf