LongEval-Retrieval: French-English Dynamic Test Collection for Continuous Web Search Evaluation

Petra Galuščáková[*]
Univ. Grenoble Alpes, CNRS,
Grenoble INP[†], LIG
Grenoble, France
petra.galuscakova@univ-grenoble-alpes.fr

Romain Deveaud[*]
Qwant
Paris, France
r.deveaud@qwant.com

Gabriela González Sáez
Univ. Grenoble Alpes, CNRS,
Grenoble INP[†], LIG
Grenoble, France

Philippe Mulhem
Univ. Grenoble Alpes, CNRS,
Grenoble INP[†], LIG
Grenoble, France

Lorraine Goeuriot
Univ. Grenoble Alpes, CNRS,
Grenoble INP[†], LIG
Grenoble, France

Florina Piroi
Research Studios Austria,
Data Science Studio
Vienna, Austria

Martin Popel
Charles University
Prague, Czechia

ABSTRACT

LongEval-Retrieval is a Web document retrieval benchmark that focuses on continuous retrieval evaluation. This test collection is intended to be used to study the temporal persistence of Information Retrieval systems and will be used as the test collection in the Longitudinal Evaluation of Model Performance Track (LongEval)[1] at CLEF 2023. This benchmark simulates an evolving information system environment – such as the one a Web search engine operates in – where the document collection, the query distribution, and relevance all move continuously, while following the Cranfield paradigm for offline evaluation. To do that, we introduce the concept of a *dynamic test collection* that is composed of successive sub-collections each representing the state of an information system at a given time step. In LongEval-Retrieval, each sub-collection contains a set of queries, documents, and soft relevance assessments built from click models. The data comes from Qwant, a privacy-preserving Web search engine that primarily focuses on the French market. LongEval-Retrieval also provides a "mirror" collection: it is initially constructed in the French language to benefit from the majority of Qwant's traffic, before being translated to English. This paper presents the creation process of LongEval-Retrieval and provides baseline runs and analysis.

[*]Both authors contributed equally to this research.

[†]Institute of Engineering Univ. Grenoble Alpes.

[1]https://clef-longeval.github.io/

CCS CONCEPTS

• **Information systems** → **Test collections**; **Relevance assessment**; **Multilingual and cross-lingual retrieval**.

KEYWORDS

web search, datasets, evaluation, cross-language retrieval

1 INTRODUCTION

The LongEval test collection is proposed to support the evaluation of commercial and open-source state-of-the-art Web information retrieval systems. This test collection is dedicated to provide a large scale evaluation and it is able to cope with the temporal evolution of real Web data. In order to conduct a proper evaluation of the search engines, we rely on the classical *Cranfield paradigm*. Instead of providing one test collection, we build a succession of test collections, each of them composed of: a set of documents from Qwant's actual index, a set of topics acquired from actual user's queries, and two sources of relevance assessments that estimate which documents are relevant for each topic. We call this setup to be a continuous evaluation, which is done using dynamic test collections.[2]

[2]This setup thus differs from the concept of continuous evaluation used by Tonon et al. [32], who refer to this term when they expand the set of relevance judgements iteratively for new retrieval approaches. Similarly, Carterette et al. [8] refers to Dynamic Test collection as collections, in which the queries and other actions such as clicks and time spent reading documents can be generated in response to the system, while we refer to Dynamic Test collection to a collection which evolves in terms of documents, queries, and relevance. Our concept of the Dynamic collection is thus similar to the concept used by Soboroff [31].

More specifically, the LongEval Web search test collection aims at evaluating IR systems with a focus on:

- the methods for connecting the predictivity and robustness of the offline and online evaluation, and
- the temporal persistence of IR systems.

The corpus provides relatively large amounts of collected, filtered and cleaned Web pages (several millions) in two languages (original French documents and their automatic English translations). We consider the collection to be a useful experimental resource for Websearch engines and Information Retrieval systems in general. Moreover, the provided high-quality translations of both queries and documents might be helpful for the researchers working on cross-language retrieval.

1.1 Use Cases

The Longeval Websearch collection relies on a large set of data (corpus of pages, queries, user interaction) provided by a commercial search engine (Qwant). It is designed to reflect the changes of the Web across time, by providing evolving document and topic sets. To our knowledge, such features do not exist in current test collections, at the scale we are focusing on.

The collection, as defined, aims at answering fundamental questions on the robustness and the stability of Web search engines against the evolution of the data. Such questions are still seen to be emerging and important for the community [16]. Regarding Web search evaluation, LongEval focuses on the following questions:

- How does the search engine behave as the collection evolves? Such a question is especially important for commercial systems, as the satisfaction of users is the core of such systems.
- When do we need to update an IR system as the collection evolves? If we are able to assess the decrease (if any) of a system on an evolving collection, we may then decide if the system needs to be updated.

More globally, as the collection is large (millions of documents, thousands of queries, large assessment sets), it is usable as a good training source for deep learning methods. Such real source of data is rare enough to be very valuable for the community. Though our main focus is Web search, answering similar questions on a robustness of the predictions is also in the interest of neural information retrieval systems trained and tested on different datasets. The collection, as defined, can also help studying in detail the robustness of such deep models against novelty (documents, queries). Such data is a must for the community as this question is still largely open for IR.

1.2 Related Work

Several collections, which contain information about Web users and their interactions were recently published, such as Sogou-QCL [38], MS MARCO [4], and Orcas [13]. These collections are large and thus provide an excellent ground for training the machine learning retrieval models. However, the evolution of the Web data is not considered in these collections in any way. Other recently published Web search collections TripClick [28] and TripJudge [3] also provide information about the publishing date of the documents. These collections focus on the medical domain as they use data

from the search engine working over the Trip medical database[3]. The provided interactions are directly mined from the log files and the collection in total contains 1.3 million pairs of query-document relevance signals collected over the years 2013-2020. Other collections, such as Robust [34] and Twitter [30] also contain information about the time at which was each document in the collection created. Such information might be used to study the changes in the collection and the behaviour of the search engines under these changes. However, collections which somehow consider an evolving environment as such are rare and, to our best knowledge, the only collection focused on similar questions is the recent TREC-COVID dataset [33].

TREC-COVID proposes an evolving corpus with topics from real questions extracted from forums. Compared to the LongEval collection, the number of documents and the number of topics is rather small (few tens of thousands documents and 45 topics). TREC-COVID is focused on studying a very specific topic, the COVID outbreak. In our case, the topics are much broader, and the queries are actual user's ones. Thus, the LongEval collection is the only large collection with up-to-date data (acquired in 2022) that exists to evaluate modern IR retrieval on evolving data. This also differentiates the LongEval collection from other cross-language collections which contain English-French language pairs. French and English were, for example, part of the early TREC [35] and then CLEF [5–7] cross-language retrieval evaluation campaigns. From the later collections, French and English were in a limited way a part of CLEF eHealth Collections [20, 21].
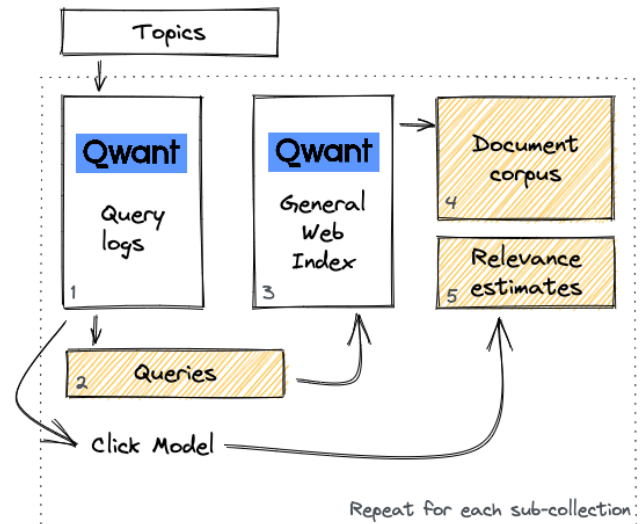
Figure 1: The general process for collecting and building the LongEval-Retrieval collection. Stroked boxes represent parts that are contained in each sub-collection.

---

[3]https://www.tripdatabase.com/

2 LONGEVAL-RETRIEVAL COLLECTION

In this section, we describe the general acquisition process of the data from the Qwant Web search engine, and the creation of different components of the collection. The overall acquisition is periodic and is recurrent over time in order to build a sequence of *sub-collections*. A sub-collection presents all the characteristics of a traditional test collection with queries, documents, and relevance judgments, except that it shares a common set of *topics* (which are not the queries themselves) with the other sub-collections. The overview of the acquisition is presented in Figure 1. It consists of:

(1) The acquisition of a set of **topics**, selected from the Web and social media. This acquisition is based on trending – yet stable in the long-term – topics and is performed only once for the entire LongEval-Retrieval collection. The selection of the set of topics is further described in Section 3.1;

(2) The selection of **search queries** related to the topics, coming from the actual queries issued by users to Qwant. We detail the query selection method in Section 3.2;

(3) The creation of **relevance estimates**. We rely here on two ways of gathering such assessments: implicitly using Click Models [11] computed from Qwant's query logs, and explicitly using manual assessments which would be gathered in the forthcoming future. Since each sub-collection can contain several thousands of queries, we will perform explicit assessments on a limited subset of manually selected queries. We present our methodology for generating relevance estimates in Section 4;

(4) The acquisition of the **document corpus** (Section 5). This corpus from Qwant is a union of: i) all the Web documents that have been displayed for each query of a sub-collection, and ii) a sizable random sample of Qwant's index. This protocol leads to a corpus that contains a mixture of relevant and non-relevant documents. The presented process handles the evolution of the Web pages, as the corpus is not only composed of URLs, but also of the Web page contents acquired at the time t.

As described, the LongEval corpus evaluates systems along time. To do that, the acquisition presented is achieved periodically, typically each month. In each time period t, we create a collection composed by the queries, assessments and documents collected during this month. The full LongEval-Retrieval dataset, composed of a sequence of collections, is thus dynamically evolving. This allows us to create and provide test collections for different time periods.

3 FROM TOPICS TO QUERIES

As described above, the first step of the data acquisition is the definition of a set of *topics*, which act as a proxy for controlling the themes expressed in the queries. Those topics are later used for filtering and selecting the Qwant actual user's *queries*. The Qwant search engine processes approximately 200,000 queries per day. These queries follow the conventional long-tail distribution. Therefore, we need to define a method for sampling intriguing and demanding queries that are likely to generate multiple user interactions. We decide to select the query topics corresponding to trending topics (i.e., topics that are popular among the queries

asked to Web search engines) in English and French. This choice supports our use case of studying Web search engines, while it is also a good choice for examining temporal changes.

DEFINITION 1. *In the context of the LongEval-Retrieval test collection, a topic is a short multi-word term – composed of either one or a few tokens – with a broad and potentially ambiguous meaning.*

The topics then serve as an entry point for selecting a different set of queries for each sub-collection. We allow the collection to have a shifting query set from one sub-collection to another in order to follow the potential drifts in the query distribution.

DEFINITION 2. *In the context of the LongEval-Retrieval test collection, a query is a multi-word chain of characters that is related to one or more topics.*

We detail how we selected the topics of LongEval-Retrieval in Section 3.1, before presenting in Section 3.2 the method that we used to define a set of queries for each topic.

3.1 Topics Selection

The list of topics has been generated once for the entire LongEval-Retrieval collection in May 2022. Topics were selected to ensure a balance between four elements: popularity, stability, generality, and diversity. We targeted long-term trends of diverse domains to avoid selecting topics that were related to sudden interest shifts so as not to bias our collection towards events. The set of topics is selected from English and French languages to ensure a multilingual validation of our findings, although a large majority of Qwant users are French. To ensure some degree of stability in the entire dynamic collection, we defined a single set of topics that is common to all sub-collections.

Topics should be popular: this criterion should ensure that there will be enough potential queries covering this topic in Qwant's query log, and in turn, enough user feedback.

Topics should be stable in time: it is important to be able to consider topics that do not evolve over time, so that we can really assess longitudinally the behavior of systems. To do that, we measure the persistence and recurrence in time on historical trending topics using Google Trends[4] and Qwant's query log.

Topics should be general enough to cover numerous queries. We sampled queries from a 1-month Qwant's query log, and kept the queries that overlapped with our candidate topics. We considered topics for which a large number of queries matched ($\geq 1,000$ queries) as general enough.

Table 1 presents the set of topics that we used to further select the queries of the sub-collections.

3.2 Queries Selection

The topics previously extracted following the methods described in the previous section are a first step towards the identification of real user queries that are issued on Qwant, which we will use for each sub-collection. Qwant answers a portion of user queries with their own technology while the rest is forwarded to a third party search engine. In order to ensure that all displayed documents are

---

[4]https://trends.google.com

	Topic	English Description
1	eau	water
2	nourriture	food
3	espace	space
4	voiture	car
5	argent	money
6	manifestation	protest
7	virus	virus
8	terre	earth
9	énergie	energy
10	police	police
11	loi	law
12	travailleurs	workers
13	guerre	war
14	invasion	invasion
15	président	president
16	pétrole	oil
17	impôts	taxes
18	votants	voters
19	fraude	fraud
20	elisabeth borne	(French prime minister, May, 2022)
21	changement climatique	climate change
22	fête du travail	(French holiday, usually associated with protests)
23	eurovision	eurovision
24	jacques perrin	(French actor)
25	regine	(French singer)
26	heartstopper	-
27	fete des meres	mother's day
28	johnny depp	-

Table 1: List of the French topics of the LongEval-Retrieval collection that were the basis for selecting the queries. The English description is informative and has not been further used to build the collection.

contained in Qwant's index, we restricted the query distribution that we consider to the queries that Qwant actually answers.

We use a simple text processing technique to map general topics to queries that are answered by Qwant. Let Q be the set of all queries that are answered by Qwant, and $\mathcal{T}$ be the set of topics we defined in Section 3.1. For each topic $t \in \mathcal{T}$, we select all the queries Q_t from Q that contain t as a sub-string (denoted by $\subseteq_{str}$):

$$Q_t = \{q | q \in Q, t \subseteq_{str} q\}$$

Then, for the full set of topics, we have:

$$Q_{\mathcal{T}} = \cup_{t \in \mathcal{T}} Q_t$$

As this filtering can lead to several tens of thousands of queries per topic when considering several days/months of query logs, we applied a top-k selection for each Q_t, noted $Q_t^{top_k}$, using which we only keep the k most frequently asked queries on Qwant for each topic. Finally, $Q_{\mathcal{T}}^{top_k}$ denotes union of $Q_t^{top_k}$ over the topics.

3.3 Queries Filtering

Following the initial query generation process, the queries undergo automatic filtering to select those with at least 10 relevance assessments (as detailed in the following section)[5]. Subsequently, a manual check is performed by a human annotator to ensure query quality. During this process, queries with similar objectives, such as

[5]It should be noted that some queries may have fewer than 10 assessments in the final collection due to further corpus filtering applied after this step.

'achat voiture' and 'acheter voiture', 'anti virus' and 'antivirus', or 'bareme impots' and 'barême impots', are merged. Queries referring to adult content are removed from the collection.

4 RELEVANCE ESTIMATES

The relevance estimates for LongEval-Retrieval are obtained through automatic collection of users' implicit feedback. However, in Section 9, we discuss the future completion of these estimates with explicit relevance assessments.

In order to be consistent with the creation of the query and document collections, we rely on user clicks as an implicit feedback to automatically infer the relevance of different documents. Qwant is committed to protecting online privacy, and does not track information such as multiple clicks, dwell times, or query reformulations. The only information logged by Qwant is the query, the displayed documents, and the rank at which a click occurred.

Raw clicks cannot be used as a signal of relevance due to their tendency to be noisy and heavily biased [18] towards the top-ranked results. Noise comes from the fact that a click does not necessarily indicate relevance, while a lack of click does not indicate irrelevance. Aggregating larger samples of query logs can remove noise, but statistical biases remain. Such biases can result from various factors, such as position bias towards top-ranked results, presentation bias towards visually appealing results, or trust bias towards results from familiar domains [24].

4.1 Click Models

To tackle this problem, we opted to *debias* the click data by estimating Click Models. A Click Model [12, 37] is the base to infer the user relevance of a document from search log data. It computes the estimates of a document's attractiveness given a query and thus it tackles the problem of using the users' interaction while avoiding sharing private data and reducing noise and bias. Many click models have been developed over the years with the goal of better modelling the clicking behaviour of users, mostly by investigating sessions of multiple clicks for a given query. In our case, since Qwant is a privacy-preserving search engine and does not track any search session, we cannot rely on the more advanced Click Models that consider multi-query and multi-click sessions.

We thus implemented the simplified version of the Dynamic Bayesian Networks (DBN) [10] with a session length of 1, which comes down to the original Cascade Model [15]. The Cascade Model assumes the following user model: users scan a Search Engine Results Page (SERP) from top to bottom, skipping non-attractive documents and clicking on what they believe will be relevant based on attractiveness. By fitting such a model, we were able to use the learned attractiveness parameter α as a soft relevance assessment. Such assessment of the relevance of a document given a query is both a) noise-free and b) unbiased with respect to the underlying user model. The α parameter of the model is then equal to the probability that a document would be attractive given a query. Maximum Likelihood Estimation (MLE) can be used to estimate α for each pair of query-document. For a given query q and a document d, the attractiveness probability is estimated as follows:

$$\alpha_{q,d} = \frac{1}{|\mathcal{S}_{q,d}|} \sum_{s \in \mathcal{S}_{q,d}} c_s(d)$$

where $\mathcal{S}_{q,d}$ is the set of all entries in Qwant's query log where document d has been displayed in a SERP above or at the rank of the clicked document, and $c_s(d)$ is a binary variable indicating where document d has been clicked in the entry s.

We estimated this parameter by using MLE over Qwant's query log filtered for the queries and the time period of each sub-collection. The attractiveness α then provides us with a surrogate for document relevance for each query. However, traditional IR metrics such as Precision, Mean Average Precision, or nDCG rely on integer relevance assessments (either binary or scaled). This makes the attractiveness probabilities as relevance estimates impractical for computing IR metrics. We thus transformed α into discrete values by reversing a function $f(g)$ [9] used for mapping a discrete relevance judgment g to a probability of relevance, based on the Discounted Cumulative Gain discount function:

$$f(g) = \frac{2^g - 1}{2^{g_{max}}}$$

where g_{max} is the maximum grade of relevance (i.e. $g_{max} = 1$ for binary relevance judgments). We simply reversed the function as follows:

$$P = \frac{2^g - 1}{2^{g_{max}}}$$
$$\Leftrightarrow P \times 2^{g_{max}} = 2^g - 1$$
$$\Leftrightarrow g = log_2(P \times 2^{g_{max}} + 1)$$

where $P = \alpha_{q,d}$ in our case.

Using Qwant's query log, we then compute both probability-based and discrete relevance estimates for all query-document pairs that we gathered using the methods previously depicted. We specifically used $g_{max} = 2$, which corresponds to:

- 0 = not relevant,
- 1 = relevant,
- 2 = highly relevant.

4.2 Post-filtering

The last step of building the relevance estimates is their post-processing according to the filtering done on the queries. As described above, the queries were manually checked and some of the queries were removed. In such case, all the relevance estimates corresponding to such queries were removed as well. Similarly, some of the queries were merged and in such cases, the corresponding estimates were also merged. If, during the merge, the assessments of the queries were having contradictory values (e.g. the document was relevant according to one query and not relevant according to the second query), these relevance estimates were excluded from the collection.

5 DOCUMENT CORPUS

The final part of the test collection is related to the corpus of documents, on which the relevant documents are selected to be retrieved for a specific query. The first step for creating the document collection is to extract from the index the content of all the documents that have been displayed in SERPs for the queries that we selected (see Section 3.2). In addition to these documents, potentially non-relevant documents are randomly sampled from Qwant's index in order to better represent the nature of a Web test collection.

5.1 Random Sampling

Two major problems can occur if we built the document collection only using the documents of the relevance assessments: 1) the collection will be heavily biased towards Qwant's ranking function, and 2) the prevalence of relevant documents will be artificially high. We alleviated this problem by sampling random documents for each original topic and introducing them into each sub-collection. For each topic $t \in \mathcal{T}$, we randomly selected at most $n = 100,000$ documents among those that matched the word tokens of t[6]; this number could be lower depending on the number of documents actually matching the topic. Such selection avoids oversimplifying the corpus and the search task, as these documents are not completely randomly picked from Qwant's index.

5.2 Cleaning Process

The collection does not only provide the url of each document in the corpus, but also cleaned versions of these documents. To do so, we first extract the text content from the websites, for which we use the internal Qwant implementations. Thus, we are able to use exactly the same representations of documents as Qwant uses for ranking the documents. We also apply an adult and spam content filter on the built collection. Even though this filtering is quite strict, adult and spam content is still very frequent in the collection.

6 ENGLISH TRANSLATIONS

Given that the vast majority of Qwant users are French speakers, the search engine's primary focus is on searching for and analyzing French queries and data. However, to make the collection more accessible to non-French speakers, we also provide English translations for both queries and documents. As a result, the collection can be used for monolingual English searches and is also well-suited for cross-lingual retrieval purposes.

Initially, all the queries and documents were created in French. To generate their English counterparts, we used a machine translation system to translate the French queries and documents into English. For the translation, we used the French-English CUBBITT (Charles University Block-Backtranslation-Improved Transformer Translation) system [25], available at the LINDAT/CLARIAH-CZ infrastructure[7]. The English–French parallel training data for CUBBITT were downloaded from WMT 2014 [8] and the monolingual data were downloaded from WMT 2018, and were filtered to only contain French and English content. Naturally, the quality of the translations of the queries, which are often very short, is much lower than the quality of the document translation. This is due to the mismatch of domain of training data which is not suitable for query translation, and also due to that the translation system has in the case of the documents much more contextual information available.

To translate the documents, we first segment all the websites into sentences using the French model implemented by Spacy[9]. We then detect the language of each sentence using Fasttext [19] and

---

[6]We used a very basic AND matching, hence selected documents had to contain all the word tokens of the topic.
[7]https://lindat.cz/services/translation
[8]http://statmt.org/wmt14/translation-task.html
[9]https://spacy.io/

only translate the sentences that are highly likely to be French. Otherwise, we skip the translation and keep the original sentence. We also exclude single-character sentences from translation, as they often lead to catastrophic errors. To ensure tractability, we limit the translation to truncated sentences of up to 500 bytes. Our experiments (see Section 8) demonstrate that this approach has little impact on the evaluation measures.

7 LONGEVAL COLLECTIONS

LongEval [1] is a shared task, which will be for the first time organized in 2023 at CLEF. The task is dedicated to evaluating the temporal persistence of Information Retrieval (IR) systems and Text Classifiers. It is motivated by recent research showing that the performance of the models drops as the test data becomes more distant, with respect to time, from the training data. This is true for classification [2, 17, 22], but also the research in information retrieval shows that deep neural network-based IR systems are dependent on the consistency between the train and test data [29]. To be able to study this, one needs several test collections created over sequential time periods, which would allow doing observations at different time stamps t, and most importantly, comparing the performance across different time stamps t and t'. We want to specifically investigate two particular scenarios: short-term persistence and long-term persistence. The short-term persistence task aims to assess the performance difference between t and t' when t' occurs right after or shortly after t. In the long-term persistence task, we aim to examine the performance difference between two t and t'', when t'' occurs several months after t (and thus $|t'' - t| > |t' - t|$).

So far, we described the creation of the collection at the single timestamp t referring to a single month. We repeat this process of collecting the data over several months and create the Train Collection t and two test collections t' and t''.

7.1 Train Collection

Train collection is collected during June 2022 and was released on the Lindat infrastructure[10]. The document corpus consists of 1,570,734 Web pages. As we would like to be able to compare the systems between t and t' or t'', we further split the queries in this Train collection into train and heldout queries. The collection consists of 672 train queries, with corresponding 9,656 assessments and 98 heldout queries with corresponding 1,420 assessments. There are thus in average 14 assessments per query. About 73% of the assessments are non-relevant (7,030 assessments on the train queries in total), 21% are relevant (2,028 assessments) and 6% are highly relevant (598 assessments).

We display several randomly selected examples of the queries and their translations in Table 2. Though the majority of the queries, as displayed, is translated reasonably, some of the translations still failed. For example, the query "cuisson gigot agneau" (meaning "cooking leg of lamb") is translated as "leg" repeated numerous times. The original French queries are relatively short – most of the queries consists of either three (42%) or two (30%) words. Only a single query consists of 7 words, and there is no longer query

than that. In Table 3[11] we show an example of highly relevant and non-relevant document.

While we publish the assessments for the train queries, we keep the assessments for the heldout queries hidden from the participants and only release them after publishing the official task results. The participants are required to submit the results for either one or both test collections and also to submit the results on the heldout queries. This setup allows us to measure the differences of the submitted systems between t and either t' or t''.

7.2 Test Collections

We provide two test collections[12]: the short-term persistence collection is collected over July 2022 and the long-term persistence collection over September 2022. Both these collections have similar sizes as the train collection: the short-term persistence collection consists of 1,593,376 documents and 882 queries and the long-term persistence collection consists of 1,081,334 documents and 923 queries. The relevance judgements will be released after publishing the official task results, and the amounts of the judgements are also similar to the Train collection, with 12,217 and 13,467 relevance judgements for the short-term and long-term tasks, respectively.

For the evaluation, we plan to apply nDCG measure to all the systems submitted to the task on the heldout data and on both test sets. This classical evaluation measure should be well suitable for evaluating Web search, for which the discount emphasises the ordering of the top results. Apart from this, we plan to calculate Relative nDCG Drop (RnD) measured by computing the difference between nDCG on the heldout queries and test collection. This measure supports the evaluation of the impact of the data changes on the system's results. Using these two evaluation measures, a system that achieves good results using nDCG, and also good results according to the RnD measure is considered to be able to cope well with the evolution over time.

8 EXPERIMENTS

We describe here preliminary tests in a way to show that the data provided is hard enough to be a good source for research purposes, with a large margin of improvement for systems. It thus allows to validate our automatic relevance assessment process to some extent. In this first step, we use two default IR systems, Terrier [23] and Anserini [36], on both French and English corpora, using their respective default BM25 parameters, with results sizes of 1000 documents. We apply French stopwords and stemming in Terrier when processing the French collection, but we do not use any specific French processing in Anserini. We also choose these systems to check both of our data formats: Anserini relies on json format and Terrier on TREC format. The collection contains the data in both these formats, what allows easy application of both Anserini and Terrier systems. Moreover, we also use the neural T5 model [27] and its implementation as a monoT5 provided by PyGaggle[13]. Following the setup by Pradeep et al. [26], we apply T5 model as a reranker to reorder top 100 documents retrieved by the Terrier system.

---

[10]http://hdl.handle.net/11234/1-5010

[11]Documents are shortened for display purposes.
[12]http://hdl.handle.net/11234/1-5139
[13]pygaggle.ai/

Query ID	French Query	English Query
q06229550	bareme impots	Taxation
q06223863	consommation eau	consumption water
q06221247	gateau aux pommes	apple cake
q06225303	offre emploi	offer of employment
q06226031	terrence malik	terrence malik
q06226042	tesla voiture	tesla car
q06223096	visiter les chateaux de la loire	visit the castles of the loire

Table 2: Examples of the train queries.

Highly Relevant Document	Non-Relevant Document
Bordeaux Airport - Merignac - Official website Re-opening of the billi terminal: all the information you need for your next trip Our billi terminal has reopened to operate some easyJet and Ryanair flights. Shops were also reopened, in particular to offer a catering service to travellers. This decision will make it possible to deploy on the entire infrastructure the material and human resources necessary for the smooth running of the journey. Important: All other airlines: concourse A Flight boarding: check your concourse and boarding gate in real time on our website or on the terminal screens when you arrive. With large crowds, think ahead of your visit: Due to the heavy summer traffic, passengers are advised to anticipate their arrival well in advance. Wings for the World News and events Published: 10 June 2022	Parking near Bordeaux Airport Merignac | Beep Valet Parking contact@beep-valet-parking.com BEEP VALET Choose your seat type Parking secure + Shuttle offered + Shuttle Airport You are brought to the airport from the car park Reservation Airport Free Parking + Shuttle Parking Available Your windscreen needs replacing? We offer parking Reservation Parking Offered Secure payment by bank CIC Benefits of your car park Merignac airport RESERVE RAPIDE GAREZ EASY TALK SEREIN Close airport Mérignac Our tarmac car park is located 3 km from Bordeaux-Mérignac airport. Take 5 minutes of transfer time in our comfortable shuttle. If needed, child boosters and baby seats are offered. Airport Shuttle

Table 3: Examples of the translated documents relevant and non-relevant for the query "aeroport bordeaux", which was translated as "airport".

The results are presented in Table 4 for the train and heldout query sets run on the Train collection and in the Table 5 for both Test collections. The first three rows present the results on the French corpus: we see that systems on all collections perform reasonably well, with P@10 close to 0.1 and NDCG almost 0.37. In all cases, the Recall@1000 is close to 0.75, which is a reasonable score supporting experimentation with reranking IR approaches. The results of both BM25 implementations are, not surprisingly, very similar and the results for the English translations are around 15% lower than on the French documents. The interesting point is that the translation does not impact dramatically the results: this validates the quality of the translations and confirms that the translated English corpus may be used with a confidence. The comparison of the results on different collections shows that the systems perform very similarly on these collections. This confirms that the heldout queries and the test collections are good representatives for the Train collection.

9 DISCUSSION

The average number of documents retrieved by the BM25 models described in the previous section for which we have assessments available is very small: in average there are only 2 or 3 assessed documents in the top 10 retrieved documents and only around 7 assessed documents in the top 100 retrieved documents. These numbers are similar for all the runs, but slightly higher for Terrier and especially when it is applied on the English data. We would like

to deal with this problem by using explicit relevance assessments from the users, using classical TREC pooling methods.

This explicit feedback will be provided for a selected subset of queries defined in Section 3 and for the pooled list of documents which will be retrieved by the systems submitted to the LongEval-Retrieval track. The relevance assessments are thus expected to be available in summer 2023. This step is especially important to be able to compare the predictions done by the click model with more precise human assessments and thus, in turn, to compare offline and online evaluation methods. It might also help to make evaluation more fair for the systems utilizing English data, as these can be expected to retrieve different sets of documents as the systems utilizing French data.

Moreover, the explicit assessments might help to deal with the noise in the implicit ones. The proposed test collection aims at representing a Web environment, with a focus on topics and documents which change over time. Given these characteristics of our test collections, we may have possible noise/error induced on the relevance assessment when we cluster several one-click user logs, in order to infer the complete user interaction. It is due to the fact that Qwant stores limited user interaction information.

Based on the recent observations from TREC [14], the selection of the queries for the pool will be crucial for the explicit assessments. While the selected queries need to be representative and also need to have enough relevant documents, the number of relevant documents still needs to be limited. If the amount of relevant

		Train				Heldout			
System	Language	MAP	P@10	NDCG	Recall	MAP	P@10	NDCG	Recall
BM25 Anserini	fr	0.1617	0.0987	0.3102	0.7368	0.1676	0.1112	0.3249	0.7573
BM25 Terrier	fr	0.1767	0.1109	0.3308	0.7695	0.1904	0.1184	0.3540	0.8119
BM25 Terrier + T5	fr	0.2175	0.1329	0.3650	0.7695	0.2223	0.1408	0.3757	0.8119
BM25 Anserini	en	0.1372	0.0887	0.2768	0.6766	0.1267	0.0878	0.2646	0.6367
BM25 Terrier	en	0.1457	0.0938	0.2893	0.7048	0.1326	0.0908	0.2689	0.6372
BM25 Terrier + T5	en	0.1832	0.1141	0.3211	0.7048	0.1576	0.1092	0.2886	0.6372

Table 4: BM25 results for Anserini and Terrier (top-1000 documents), on the Train collection and the *train* and *heldout* query sets of LongEval. Terrier BM25 results are presented without (BM25 Terrier) and with (BM25 Terrier + T5) reranking applied on the top 100 retrieved documents performed by the T5 model.

		Short-term				Long-term			
System	Language	MAP	P@10	NDCG	Recall	MAP	P@10	NDCG	Recall
BM25 Anserini	fr	0.1811	0.1130	0.3292	0.7426	0.1784	0.1207	0.3338	0.7581
BM25 Terrier	fr	0.2027	0.1234	0.3541	0.7781	0.1936	0.1336	0.3526	0.7928
BM25 Terrier + T5	fr	0.2209	0.1387	0.3717	0.7781	0.2207	0.1472	0.3801	0.7928
BM25 Anserini	en	0.1437	0.0907	0.2783	0.6630	0.1429	0.0984	0.2883	0.6921
BM25 Terrier	en	0.1510	0.0948	0.2873	0.6811	0.1536	0.1034	0.2992	0.7084
BM25 Terrier + T5	en	0.1863	0.1145	0.3202	0.6811	0.1936	0.1271	0.3347	0.7084

Table 5: BM25 results for Anserini (top-1000 documents), on the Test collections for the *short-term* and the *long-term* LongEval tasks. Terrier BM25 results are presented without (BM25 Terrier) and with (BM25 Terrier + T5) reranking applied on the top 100 retrieved documents performed by the T5 model.

documents is too high, it might not be possible to easily distinguish the quality of the systems, which we would like to compare.

10 PRIVACY AND ETHICAL CONSIDERATIONS

The LongEval-Retrieval collection is composed of queries that have been issued by real users to the Qwant search engine, thus raising concerns related to the privacy of these users especially amidst of scandals such as the AOL query log release[14]. Qwant is a privacy-preserving search engine by design: all user queries are entirely anonymised, and subsequent queries cannot be linked together to form search sessions or a search history. We also do not provide any additional contextual information apart from the text of the query. Moreover, the Click Model that we use for computing relevance estimates aggregates interactions over several instances of the same queries, thus limiting potentially specific and personal queries.

Although the privacy-preserving nature of Qwant limited the leaking of user information, a risk still exists that some queries could contain sensitive information, although this is very unlikely. In the event of such a leak happening, we will remove the associated query and provide an update to the collection.

11 CONCLUSION

This paper describes the LongEval-Retrieval test collection, which is designed specifically for evaluating information retrieval systems with evolving data. The collection is constructed from Qwant data, including Web documents and real user input. The described

data collecting and processing methods are able to cope with large amounts of data, by generating automatic queries and their relevance estimates from query logs. We create a corpus which integrates SERP results as well as "background" documents from Qwant's index. The corpus is initially in French, and LongEval-Retrieval also provides English translations of the documents and queries.

We describe in detail the process of creating the collection and also the published Train and Test collections which use this process. We also provide initial experiments which confirm the usability of these collections and provide an initial baseline. In the future, we plan to integrate manual assessments into the collection, so that we can compare manual and automatic assessments in a large scale. LongEval-Retrieval is currently able to cope with queries evolution, according to a fixed set of initial topics, but in the future, we want to extend our proposal so that we also handle the evolution of topics.

Regarding the initial objectives of LongEval-Retrieval, we aim to connect the predictivity and robustness of the offline and online evaluation approaches. Using the proposed evaluation measures, we will be able to explore the temporal persistence of IR systems across several corpora that evolve, compared to a reference one. By doing so, the collection will provide a more comprehensive and accurate evaluation of IR systems in real-world settings, where the data is constantly changing over time.

---

[14]https://en.wikipedia.org/wiki/AOL_search_log_release

REFERENCES

[1] Rabab Alkhalifa, Iman Bilal, Hsuvas Borkakoty, Jose Camacho-Collados, Romain Deveaud, Alaa El-Ebshihy, Luis Espinosa-Anke, Gabriela Gonzalez-Saez, Petra Galuščáková, Lorraine Goeuriot, et al. 2023. LongEval: Longitudinal Evaluation of Model Performance at CLEF 2023. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Proceedings, Part III*. Springer, 499–505.

[2] Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. 2022. Building for Tomorrow: Assessing the Temporal Persistence of Text Classifiers. *arXiv preprint arXiv:2205.05435* (2022).

[3] Sophia Althammer, Sebastian Hofstätter, Suzan Verberne, and Allan Hanbury. 2022. TripJudge: A Relevance Judgement Test Collection for TripClick Health Retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. ACM.

[4] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs.CL]

[5] Martin Braschler. 2001. CLEF 2000 — Overview of Results. In *Cross-Language Information Retrieval and Evaluation*. 89–101.

[6] Martin Braschler. 2002. CLEF 2001 — Overview of Results. In *Evaluation of Cross-Language Information Retrieval Systems*. Springer, 9–26.

[7] Martin Braschler. 2003. CLEF 2002 — Overview of Results. In *Advances in Cross-Language Information Retrieval*. 9–27.

[8] Ben Carterette, Ashraf Bah Rabiou, and Mustafa Zengin. 2015. Dynamic Test Collections for Retrieval Evaluation. *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (2015).

[9] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09)*. ACM, 621–630.

[10] Olivier Chapelle and Ya Zhang. 2009. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web (WWW '09)*. ACM, 1–10.

[11] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. Click Models for Web Search. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 7, 3 (July 2015), 1–115.

[12] Aleksandr Chuklin, Pavel Serdyukov, and Maarten De Rijke. 2013. Click model-based information retrieval metrics. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 493–502.

[13] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 18 Million Clicked Query-Document Pairs for Analyzing Search. arXiv:2006.05324 [cs.IR]

[14] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Ellen M. Voorhees, and Ian Soboroff. 2021. TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. ACM, 2369–2375.

[15] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, 87–94.

[16] Nicola Ferro, Norbert Fuhr, Gregory Grefenstette, Joseph A. Konstan, Pablo Castells, Elizabeth M. Daly, Thierry Declerck, Michael D. Ekstrand, Werner Geyer, Julio Gonzalo, Tsvi Kuflik, Krister Lindén, Bernardo Magnini, Jian-Yun Nie, Raffaele Perego, Bracha Shapira, Ian Soboroff, Nava Tintarev, Karin Verspoor, Martijn C. Willemsen, and Justin Zobel. 2018. From Evaluating to Forecasting Performance: How to Turn Information Retrieval, Natural Language Processing and Recommender Systems into Predictive Sciences (Dagstuhl Perspectives Workshop 17442). *Dagstuhl Manifestos* 7, 1 (2018), 96–139.

[17] Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences* 10, 12 (2020), 4180.

[18] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*. ACM, 154–161.

[19] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759* (2016).

[20] Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, João Palotti, and Guido Zuccon. 2016. Overview of the CLEF eHealth Evaluation Lab 2016. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer, 255–266.

[21] Liadh Kelly, Hanna Suominen, Lorraine Goeuriot, Mariana Neves, Evangelos Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker, Guido Zuccon, Harrisen Scells, and João Palotti. 2019. Overview of the CLEF eHealth Evaluation Lab 2019. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer, 322–339.

[22] Jan Lukes and Anders Søgaard. 2018. Sentiment analysis under temporal shift. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*. 65–71.

[23] Craig Macdonald, Richard McCreadie, Rodrygo LT Santos, and Iadh Ounis. 2012. From puppy to maturity: Experiences in developing Terrier. *Proc. of OSIR at SIGIR* (2012), 60–63.

[24] Harrie Oosterhuis. 2021. Learning from User Interactions with Rankings: A Unification of the Field. *SIGIR Forum* 54, 2, Article 16 (aug 2021).

[25] Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications* 11, 4381 (2020), 1–15.

[26] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. arXiv:2101.05667 [cs.IR]

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]

[28] Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. 2021. TripClick: The Log Files of a Large Health Web Search Engine. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2507–2513.

[29] Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qifei Wu, Yuchen Ding, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2022. arxiv:2204.12755. A Thorough Examination on Zero-shot Dense Retrieval.

[30] Royal Sequiera and Jimmy Lin. 2017. Finally, a Downloadable Test Collection of Tweets. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1225–1228.

[31] Ian Soboroff. 2006. Dynamic Test Collections: Measuring Search Effectiveness on the Live Web. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, 276–283.

[32] Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. 2015. Pooling-Based Continuous Evaluation of Information Retrieval Systems. *Inf. Retr.* 18, 5 (oct 2015), 445–472.

[33] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. *SIGIR Forum* 54, 1, Article 1 (feb 2021).

[34] Ellen M Voorhees. 2006. The TREC 2005 robust track. In *ACM SIGIR Forum*, Vol. 40. ACM, 41–48.

[35] Ellen M. Voorhees and Donna K. Harman. 1999. Overview of the seventh text retrieval conference (trec-7).

[36] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, 1253–1256.

[37] Yuchen Zhang, Weizhu Chen, Dong Wang, and Qiang Yang. 2011. User-click modeling for understanding and predicting search-behavior. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1388–1396.

[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 *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 1117–1120.