Corpora Performance Prediction

Andrew Parry[1], Jan Heinrich Merker[2], Simon Ruth[3], Maik Fröbe[2] and Harrisen Scells[4]

[1]*University of Glasgow*
[2]*Friedrich-Schiller-Universität Jena*
[3]*Universität Kassel*
[4]*University of Tübingen*

Abstract

Retrieval corpora are usually created to answer specific types of information needs. For example, medical search engines collect medical documents to support medical queries. Existing query performance prediction (QPP) methods aim to predict how effectively individual queries will be served by a search system. We instead focus on the ability of corpora to serve different types of queries. We propose the task of corpora performance prediction (CPP) to automatically predict, without relevance judgments, for what kinds of queries given retrieval corpora will be effective. Given a set of corpora, queries, and QPP methods, we obtain corpus "signatures" by measuring the performance of each query on each corpus with all predictors. Analyzing those corpus signatures, we aim to answer questions such as what search tasks are supported on each corpus? We realize our CPP concept using 250k queries from the Archive Query Log on five corpora (MS MARCO passage, a subsample of MS MARCO, Touché, NFCorpus, and Cranfield) with all predictors from the QPPTK framework. Our experiments show that the QPP methods yield helpful corpus performance predictions that characterize our corpora.

Keywords

Query Performance Prediction, Retrievability, Query Logs

1. Introduction and Background

Unlike query performance prediction (QPP), which aims to answer fine-grained questions about individual queries, we propose corpus performance prediction (CPP) to answer more general, holistic questions about an entire corpus. In general, QPP predicts how effective the results of a query will be either pre-retrieval or post-retrieval [1]. Thereby, QPP supports diverse tasks in retrieval systems such as query suggestion [2], query routing [3], query rewriting [4]. Following this spirit, our CPP task aims to use QPP to fingerprint complete corpora. Our idea is that a large set of diverse queries annotated with a taxonomic class helps to make sense of QPP predictions and to easily understand for which scenarios corpora might be used. For instance, if many queries from one class are predicted to perform well, this corpus might allow the building of a class-specific search engine, whereas corpora predicted to perform uniformly on all classes might serve for general-purpose search engines. Thereby, CPP aims to answer holistic questions about an entire corpus. CPP is only possible when a large set of queries is grouped by class. We use the Archive Query Log (AQL) [5] as it contains a large set of queries and the search provider corresponding to the search engine to which each query was submitted. Overall, the AQL has 64 million unique queries for 550 search providers. We use the search provider as the class.

Where QPP makes pair-wise comparisons between topics; we instead aggregate QPPs over a fixed set of queries that do not originate from the test collection corresponding to the corpus. Such a process is favorable as, with a large sample size, the estimation of a population mean is reduced. This allows for an unsupervised approach to meaningfully compare corpora. In probing corpora of varying size and class specificity, we explore a new way to compare corpora in terms of query effectiveness as opposed to other textual similarities. Both as a diagnostic tool and potentially as an intervention in a broader search system, CPP may provide a suitable heuristic to complement QPP. With such goals in mind, we apply several common QPP heuristics on a large scale. We run experiments on five corpora using QPPTK [6] against PyTerrier [7] indices.

QPP++ 2025: Query Performance Prediction and its Applications in the Era of Large Language Models, co-located with the 47th European Conference on Information Retrieval (ECIR 2025), April 6–10, Lucca, Italy

2. Methodology and Experimental Setup

Our experiments are intended as a preliminary setup to collect feedback during the workshop.

2.1. Corpus Performance Prediction

The main idea behind corpus performance prediction (CPP) is to use a large, fixed set of queries and a query performance predictor to probe a corpus. Formally, given a set of queries $\mathbb{Q}$ each belonging to a class $q \in \mathbb{Q}_{c_i}$, a query performance predictor f, and a document corpus $\mathbb{C}$, the objective is to aggregate the query performance predictions over each class of queries: $CPP(c_i) = \frac{1}{|\mathbb{Q}_{c_i}|} \sum_{q \in \mathbb{Q}_{c_i}} f(q, \mathbb{C})$. The corpus performance values of each class of queries can then be used to analyze differences between corpora. For this, we plot the distribution of QPP values over each class. We use 12 QPPs from QPPTK in their dockerized version [8] from TIRA/TIREx [9, 10]: *(max/avg)-IDF* [11], SCQ, (max/avg)*SCQ* [11], *var, (max/avg)-var* [11], *weighted information gain (WIG)* [12], *normalized query commitment (NQC)* [13], *score magnitude and variance (SMV)* [14], and *clarity* [15].

2.2. Data

We start our preliminary experiments with a sample of 20,000 queries from the Archive Query Log (AQL) [5] to probe information retrieval corpora (Table 1 shows examples from our sample). The corpora we investigate include the (1) MS MARCO passage collection, (2) a subsample of the MS MARCO passage collection with all documents retrieved within the top-100 results by any run submitted to the 2019/2020 Deep Learning tracks [16, 17] (this reduces the corpus size to ca. 60 000 documents while still allowing for reliable evaluation [18]; our assumption is that QPPs should identify that most non-Deep Learning queries should not work for this subsample), (3) the Touché subset of BEIR [19, 20] as an argument retrieval collection; (4) NFCorpus [21], a medical IR collection; and (5) the original Cranfield collection [22] which we include as it contains only 1 400 documents so that we expect to see that most domains can not be served from this tiny collection. The diversity of these five corpora should make meaningful differences apparent through probing.

Query Sampling A unique feature of the Archive Query Log (AQL) is that it consists of the actual search engine result pages (SERPs) of real search engines (or generally any search provider). Using the presence or absence of search results on each search engine result page as an indicator of "retrievability",[1] we first partition the queries of the AQL into retrievable and non-retrievable queries. From the 15 most popular search providers of the AQL,[2] we then randomly sample up to 10,000 retrievable and up to 10,000 non-retrievable queries per provider. The 15 search providers are then used as classes for corpus performance prediction. These classes include eight general web search engines (360, Baidu, Bing, Google, Naver, Sogou, Yahoo, Yandex), three online shops (AliExpress, Amazon, eBay), two media portals (IMDb, YouTube), a code collaboration platform (GitHub), and a microblogging website (Weibo).

Measures We use Kendall's τ to compute agreement between predictors, considering both the query-level and class-level agreement. Given that a class can only be represented by a finite number of queries, how does the sample size affect domain comparisons? Thus, we take sub-samples of classes to compute, under permutation, the probability that two classes invert in order. This can be seen as the stability of comparing domains and corpora. For a given fraction t being the portion of each class, we take N samples and compute a query performance prediction applying a predictor f on that sample. Within each dataset, we rank classes $\mathbb{Q} \in \mathcal{Q}$, where $\mathcal{Q}$ is the set of all classes, and observe if their comparison has swapped and store these values in an array $B \in \mathbb{Z}^{|\mathcal{Q}| \times |\mathcal{Q}|}$, i.e., for two classes $\mathbb{Q}_{c_1}, \mathbb{Q}_{c_2}$ if $f(\mathbb{Q}_{c_1}) > f(\mathbb{Q}_{c_2})$[3], $B_{c_1,c_2} = 1$. After all comparisons are made, the probability of a comparison inverting is thus $\frac{\min(B_{c_1,c_2}, B_{c_2,c_1})}{N}$. When the probability of two classes swapping under different permutations is high and their average QPP scores are far apart, this suggests comparison

[1] We assume that if a SERP returns no results, the query is not retrievable.

[2] As per Alexa rank, reflecting global popularity of websites, we use its latest ranking before its discontinuation in 2022.

[3] Where $f(\mathbb{Q}_{c_i})$ denotes the mean of the element-wise application of f over $q \in \mathbb{Q}_{c_i}$.

Table 1
Examples of queries from a selection of ten random providers from the AQL.

Provider	Query
360	郭飞雄 让保罗贝尔蒙多 微信上线抗疫状态 英语培训 何为摄影
amazon	humankind heritage edition playstation Levi's Women's Jeans るろうに black Adam ropa leselupe usb
baidu	cycle怎么读的 "Beno卯t Gilson" "tortsfamous.com" feb时空裂缝 16mn无缝钢管
bing	Raebareli, Uttar Pradesh wikipedia Willy de Paula Faria community renewal services chicago Ubuntu Foundation Marco Valério Messala Corvino
github	topic:minnan org:kfcd topic:fortran org:boostorg sado topic:jenkins org:lavabit topic:filesystem fork:true
google	"Entoprocta" "Darwin Deez" "Constellations" "Csereüvegek" site:hu.wikipedia.org "Samarqand Restaurant" -wikipedia "Armas e equipamentos da Guerra Russo-Ucraniana" -wikipedia
naver	바나다 알루미늄 블루투스 삼각대 셀카봉, WS-SQB641(화이트) 후기 hijrah 힐로 스테인레스 싱크롤 선반 20롤 대형, 블랙 후기 sumer 위성인터넷
yahoo	ISSN "0340-1707" payless All Size Waste Dumpsters Calgary wichita craiglists what causes vertigo in older adults belvedere palace vienna
yandex	абвгдейка телепередача 1975 Юрий Хой - Первая любовь полярные ночи милые обманщицы сериал yeralash tv periodic
youtube	deshawn #Поле_Таро DOC JAMIESON bonde das minas 15-7830

instability. To compute similarity of two corpora with respect to a reference query set, we define a representation of a corpus $\mathbb{C}$ for a predictor f and reference query set $\mathbb{Q}$ as a vector $v \in \mathbb{R}^{|\mathbb{Q}|}$ where each $v_i = f(q, \mathbb{C}), q \in \mathbb{Q}$. We compute corpus similarity as cosine similarity between QPP representations, i.e., $\mathrm{sim}(u, v) = \frac{u \cdot v}{|u||v|}$.

Table 2
Kendall's τ correlation between the NQC values of different corpora, across all search providers.

	Touche	MS MARCO	Subsample	NF Corpus	Cranfield
Touche	-	0.3225	0.4627	0.2313	0.0977
MS MARCO	-	-	0.2578	0.0845	0.0021
Subsample	-	-	-	0.2925	0.1361
NF Corpus	-	-	-	-	0.3574
Cranfield	-	-	-	-	-

3. Analysis

We describe the results of our experiments on the five corpora with the 12 QPP methods.

3.1. Comparing Corpora by Class Performance

Figure 1 illustrates three different QPP methods used to probe the five corpora. The avg-idf highlights which corpora are best at supporting which queries. In general, Cranfield and NF Corpus obtain very low avg-idf values; while MS MARCO, perhaps the most diverse of the five corpora, obtains the highest avg-idf values. The SCQ figure highlights these differences even more, while the NQC figure highlights the lower extremes of corpora that are unable to support certain queries (i.e., multilingual).

Looking at NQC in more detail, Table 2 provides correlations between the different corpora. Interestingly, the most strongly correlated corpora under NQC are Touché and the MS MARCO Subsample, while the next most strongly correlated are Cranfield and NFCorpus. The likely reason for this correlation is the similarity of their documents; the Cranfield collection is composed of research abstracts. Similarly, the NFCorpus is composed of medical abstracts; ultimately, the language used and topics covered will be similar under statistical representations, most likely leading to this correlation. The weakest correlation is between MS MARCO and Cranfield, most likely due to the very small size of Cranfield and differing content. Table 3 further breaks these results down into the correlations within each search provider. The strong correlations between Touché and the MS MARCO Subsample on Chinese corpora like 360, Sogou, and Weibo are most likely the reason for the strong correlation. This result highlights the similarities between these two corpora. Meanwhile, comparing MS MARCO Passage to Cranfield, the weakest correlations are on web search providers like Bing, Google, Yahoo, and Yandex, while the strongest correlations are on Sogou, Weibo, and Baidu. The relatively stronger correlations on these search providers may be due to them being outliers. We leave this investigation for future work.

3.2. Comparing Corpora at a Query Level

Observe in Figure 2 the comparison of QPP values across different datasets on our reference query log. It is interesting to notice that individual topics exhibit low agreement, much like query-level comparisons in Cranfield evaluation. This can partially be attributed to lexical mismatch as topics have not been designed for a particular collection. Nevertheless, aggregating topics improves the discrimination of classes with higher agreement, similar to systems in standard evaluation. In particular, compared to NFCorpus, the agreement between Touché and MSMARCO is low at a query level but improves largely at a class level. We consider that a helpful measure of corpus similarity may be how similar they are by their ability to serve a diverse reference set of queries. Thus, as outlined in Section 2.2, we measure cosine similarity between QPP representations with similarity by avg-IDF presented in Table 4. Interestingly, as seen in terms of correlation at a group level (Table 3), at a corpus level, MSMARCO is often more similar to other corpora than a subsample of itself. This may be due to QPP measures ultimately measuring the effect of a corpus in serving an information need. Thus, smaller corpora or constrained corpora exhibit similar effects.

Due to the nature of idf calculation, using out-of-class topics can lead to 0 values when a word is out-of-vocabulary; hence, several topics are grouped together. Also, due to the implementation of idf in QPPtk, several values are squashed due to a logarithm transformation; these observations are intrinsic to pre-retrieval QPP methods as opposed to our particular approach.

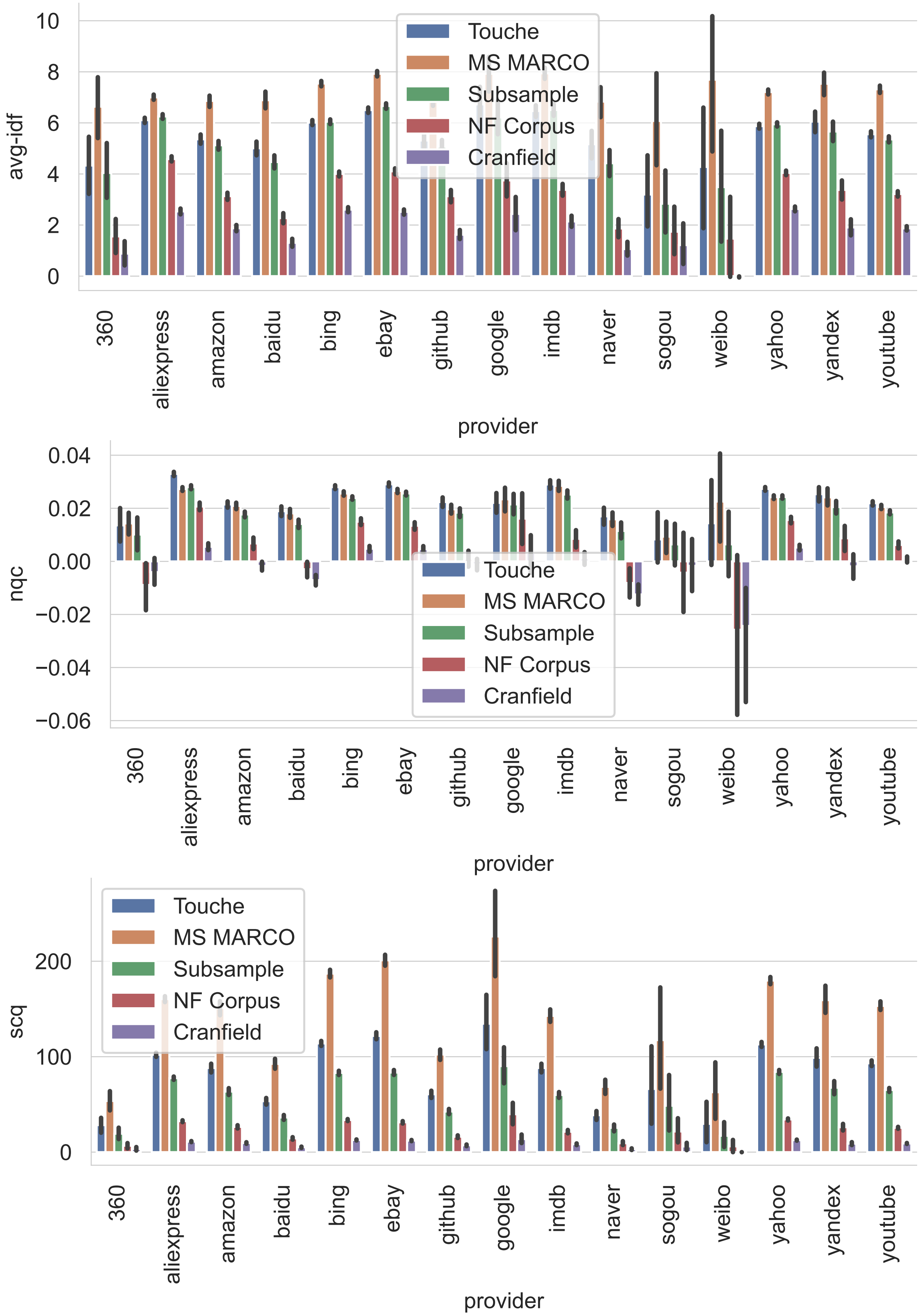


Figure 1: Distribution of AvgIDF (top), NQC (middle), and SCQ (bottom) over the three corpora. Each bar corresponds to the average value across all queries, with errors bars indicating 95% confidence interval.

Table 3
Kendall's τ correlation between the NQC values of different corpora, grouped by search providers.

		Touche	MS MARCO	Subsample	NF Corpus	Cranfield
provider						
360	Touche	-	0.3416	0.6867	0.4459	0.2210
	MS MARCO	-	-	0.2955	0.2265	0.0447
	Subsample	-	-	-	0.4661	0.2788
	NF Corpus	-	-	-	-	0.3989
	Cranfield	-	-	-	-	-
aliexpress	Touche	-	0.1249	0.3326	0.1270	0.0353
	MS MARCO	-	-	0.0853	-0.0343	-0.0470
	Subsample	-	-	-	0.1856	0.0660
	NF Corpus	-	-	-	-	0.2147
	Cranfield	-	-	-	-	-
amazon	Touche	-	0.4172	0.5539	0.3483	0.1679
	MS MARCO	-	-	0.3710	0.2361	0.1193
	Subsample	-	-	-	0.4318	0.2209
	NF Corpus	-	-	-	-	0.3921
	Cranfield	-	-	-	-	-
baidu	Touche	-	0.5011	0.5932	0.3641	0.1886
	MS MARCO	-	-	0.4314	0.2172	0.1293
	Subsample	-	-	-	0.4275	0.2552
	NF Corpus	-	-	-	-	0.3733
	Cranfield	-	-	-	-	-
bing	Touche	-	0.2949	0.4324	0.1710	0.0400
	MS MARCO	-	-	0.2462	0.0367	-0.0545
	Subsample	-	-	-	0.2180	0.0481
	NF Corpus	-	-	-	-	0.3458
	Cranfield	-	-	-	-	-
ebay	Touche	-	0.2337	0.3303	0.0980	0.0212
	MS MARCO	-	-	0.1581	-0.0223	-0.0834
	Subsample	-	-	-	0.2194	0.1132
	NF Corpus	-	-	-	-	0.3643
	Cranfield	-	-	-	-	-
github	Touche	-	0.3648	0.5179	0.3200	0.2413
	MS MARCO	-	-	0.3026	0.1500	0.1012
	Subsample	-	-	-	0.3579	0.2692
	NF Corpus	-	-	-	-	0.3954
	Cranfield	-	-	-	-	-
google	Touche	-	0.3434	0.4828	0.2420	0.0166
	MS MARCO	-	-	0.1964	0.1030	-0.0534
	Subsample	-	-	-	0.2067	-0.0662
	NF Corpus	-	-	-	-	0.3658
	Cranfield	-	-	-	-	-
imdb	Touche	-	0.2904	0.3728	0.1521	0.0949
	MS MARCO	-	-	0.1740	-0.0242	-0.0889
	Subsample	-	-	-	0.1729	0.0684
	NF Corpus	-	-	-	-	0.3722
	Cranfield	-	-	-	-	-
naver	Touche	-	0.4503	0.5588	0.3320	0.1040
	MS MARCO	-	-	0.3480	0.1421	0.0151
	Subsample	-	-	-	0.4015	0.2047
	NF Corpus	-	-	-	-	0.3584
	Cranfield	-	-	-	-	-
sogou	Touche	-	0.5029	0.8293	0.5411	0.5317
	MS MARCO	-	-	0.4383	0.2718	0.3775
	Subsample	-	-	-	0.5929	0.6189
	NF Corpus	-	-	-	-	0.3616
	Cranfield	-	-	-	-	-
weibo	Touche	-	0.4410	0.7306	0.2433	0.2533
	MS MARCO	-	-	0.2442	0.1684	0.1899
	Subsample	-	-	-	0.3785	0.2247
	NF Corpus	-	-	-	-	0.4432
	Cranfield	-	-	-	-	-
yahoo	Touche	-	0.2730	0.4360	0.1912	0.0315
	MS MARCO	-	-	0.2093	0.0360	-0.0689
	Subsample	-	-	-	0.2316	0.0531
	NF Corpus	-	-	-	-	0.3060
	Cranfield	-	-	-	-	-
yandex	Touche	-	0.3584	0.4690	0.1750	0.0184
	MS MARCO	-	-	0.2375	0.0483	-0.0255
	Subsample	-	-	-	0.2566	0.0718
	NF Corpus	-	-	-	-	0.3856
	Cranfield	-	-	-	-	-
youtube	Touche	-	0.4196	0.5532	0.2953	0.1646
	MS MARCO	-	-	0.3350	0.1454	0.0713
	Subsample	-	-	-	0.3584	0.2041
	NF Corpus	-	-	-	-	0.3988
	Cranfield	-	-	-	-	-

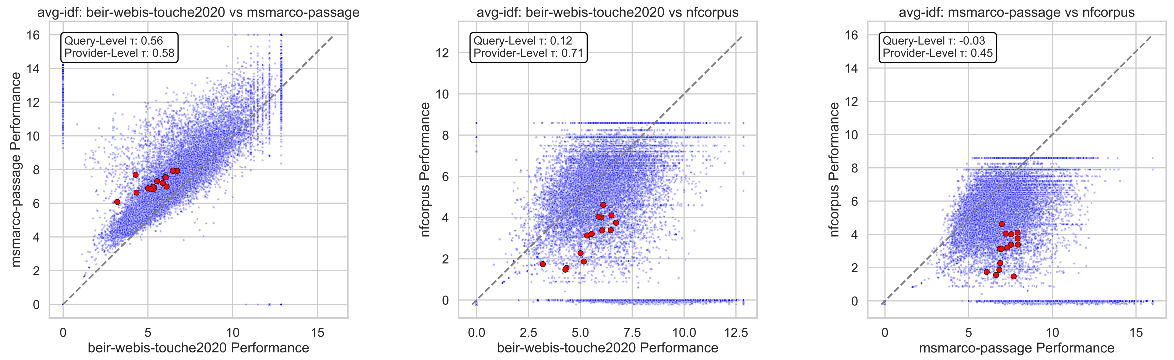
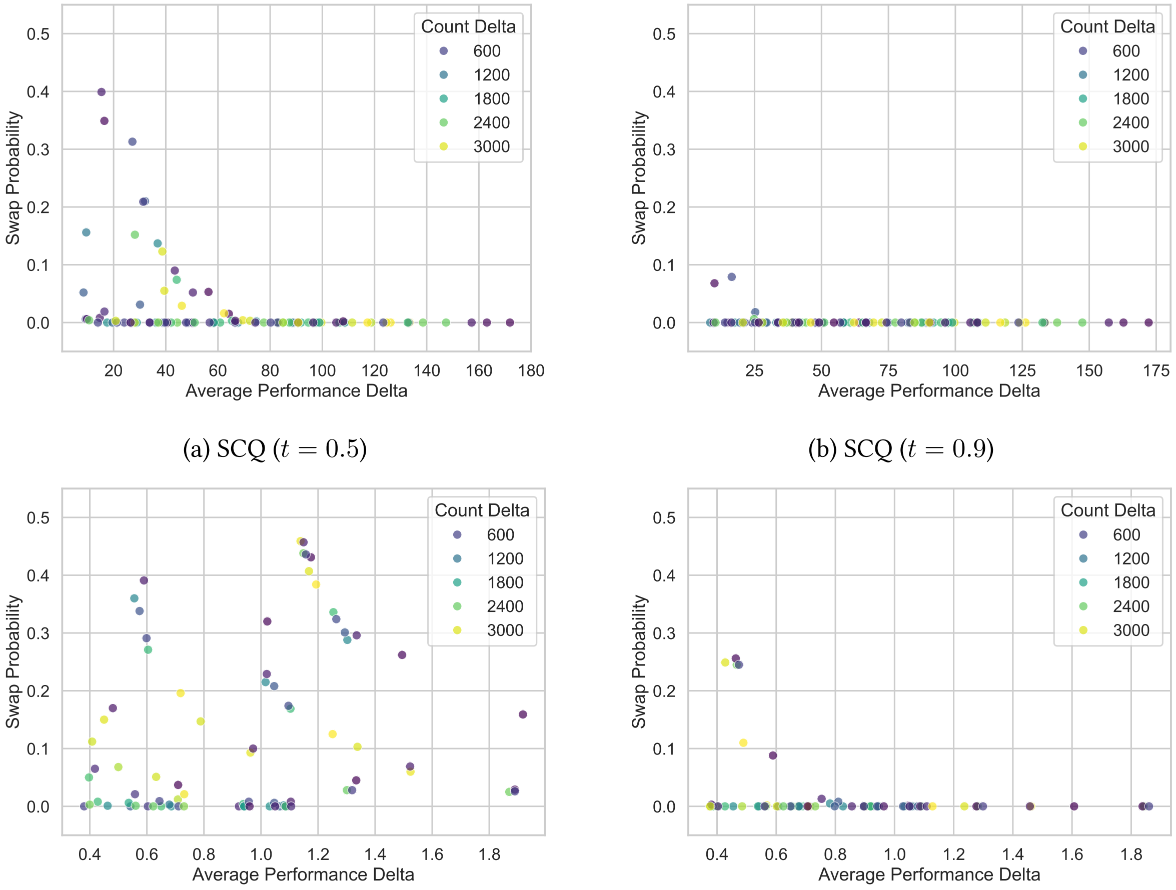


Figure 2: Query (Blue) and Provider (Red) level correlation comparing different datasets in terms of avg-IDF. Provider performance is the aggregate of the effectiveness of all constituent queries.

Table 4
Cosine Similarity of Corpora measured by QPP representations over the reference query log applying avg-IDF.

Dataset	MSMARCO	Subsample	Touche	NFCorpus	Cranfield
MSMARCO	–	–	–	–	–
Subsample	0.866	–	–	–	–
Touche	0.904	0.921	–	–	–
NFCorpus	0.700	0.804	0.746	–	–
Cranfield	0.545	0.624	0.572	0.692	–



(a) SCQ ($t = 0.5$)

(b) SCQ ($t = 0.9$)

(c) avg-IDF ($t = 0.5$)

(d) avg-IDF ($t = 0.9$)

Figure 3: The stability of provider comparisons across MSMARCO Passage using different QPP methods. t indicates the fraction of the total queries sampled.

3.3. The Robustness of class Comparisons

In Figure 3, we use sampling to simulate variable-size classes of queries, assessing the discriminative power of smaller query logs as outlined in Section 2.2. We can then compare class rankings to select a corpus or, more generally, compare corpora. Contrasting a large permutation ($t = 0.9$) with a small one ($t = 0.5$), stability improves with greater sampling, as one would expect following the hypotheses of Lesk and Salton [23] and empirical evidence of Voorhees [24] in the variations and stability of topics. Where we observe some correlation attributable to a lack of documents that can serve a query, there is no trend in the effect of class size comparisons, as shown by the Δ in class size ($|\mathbb{Q}_{c_1}| - |\mathbb{Q}_{c_2}|$) in comparisons. That is to say, we can often make strong comparisons between a small and large class, a helpful property depending on the provenance of one's query log. We observe that depending on the measure, the required sample size for stable comparison of query classes would appear to vary; however, unlike Cranfield collections, QPP can be performed without human intervention, and thus, that concern is reduced in this setting.

From the inspection of unstable class comparisons, multilingual classes tend to be more unstable, which is not unexpected given largely monolingual English corpora and QPP methods ultimately validated on English corpora. Future work could consider how neural approaches [25] may be more robust in such comparisons in QPP as they may overcome gaps in vocabulary in standard statistical methods.

4. Conclusion

We have proposed corpus performance prediction (CPP) as a framework that uses QPP measures on a large scale to make unsupervised comparisons of retrieval corpora. We validated CPP with several common English corpora of varying and diverse sizes and the Archive Query Log from which we extracted queries grouped by domain. We show that through larger samples, much like Cranfield evaluation, we can improve the stability of domain comparisons, which could assist both in diagnostics and corpora selection. As future work, we intend to scale our experiments to more corpora and to more AQL domains and queries. Other directions could be to incorporate more accurate performance predictors, e.g., with large language model relevance assessors, as they can likely produce much more accurate predictions but then only for a small representative set of queries.

References

[1] B. He, I. Ounis, Query performance prediction, Inf. Syst. 31 (2006) 585–594. URL: https://doi.org/10.1016/j.is.2005.11.003. doi:10.1016/J.IS.2005.11.003.

[2] S. Bhatia, D. Majumdar, P. Mitra, Query suggestions in the absence of query logs, in: W. Ma, J. Nie, R. Baeza-Yates, T. Chua, W. B. Croft (Eds.), Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011, ACM, 2011, pp. 795–804. URL: https://doi.org/10.1145/2009916.2010023. doi:10.1145/2009916.2010023.

[3] S. Sarnikar, Z. Zhang, J. L. Zhao, Query-performance prediction for effective query routing in domain-specific repositories, J. Assoc. Inf. Sci. Technol. 65 (2014) 1597–1614. URL: https://doi.org/10.1002/asi.23072. doi:10.1002/ASI.23072.

[4] G. Kumaran, V. R. Carvalho, Reducing long queries using query quality predictors, in: J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, J. Zobel (Eds.), Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009, ACM, 2009, pp. 564–571. URL: https://doi.org/10.1145/1571941.1572038. doi:10.1145/1571941.1572038.

[5] J. H. Reimer, S. Schmidt, M. Fröbe, L. Gienapp, H. Scells, B. Stein, M. Hagen, M. Potthast, The archive query log: Mining millions of search result pages of hundreds of search engines from 25

years of web archives, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 2848–2860. URL: https://doi.org/10.1145/3539618.3591890. doi:10.1145/3539618.3591890.

[6] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer, An enhanced evaluation framework for query performance prediction, in: D. Hiemstra, M. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I, volume 12656 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 115–129. URL: https://doi.org/10.1007/978-3-030-72113-8_8. doi:10.1007/978-3-030-72113-8\\_8.

[7] C. Macdonald, N. Tonellotto, S. MacAvaney, I. Ounis, Pyterrier: Declarative experimentation in python from BM25 to dense retrieval, in: G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, H. Tong (Eds.), CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021, ACM, 2021, pp. 4526–4533. URL: https://doi.org/10.1145/3459637.3482013. doi:10.1145/3459637.3482013.

[8] O. Zendel, M. Fröbe, G. Faggioli, QPPTK@TIREx: Simplified query performance prediction for ad-hoc retrieval experiments, in: S. M. Farzana, M. Fröbe, G. Hendriksen, M. Granitzer, D. Hiemstra, M. Potthast, S. Zerhoudi (Eds.), Proceedings of WOWS@ECIR 2024, volume 3689 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 50–62. URL: https://ceur-ws.org/Vol-3689/WOWS_2024_paper_6.pdf.

[9] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous integration for reproducible shared tasks with tira.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III, volume 13982 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 236–241. URL: https://doi.org/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6\\_20.

[10] M. Fröbe, J. H. Reimer, S. MacAvaney, N. Deckers, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, The information retrieval experiment platform, in: M. Leyer, J. Wichmann (Eds.), Lernen, Wissen, Daten, Analysen (LWDA) Conference Proceedings, Marburg, Germany, October 9-11, 2023, volume 3630 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 175–178. URL: https://ceur-ws.org/Vol-3630/LWDA2023-paper16.pdf.

[11] Y. Zhao, F. Scholer, Y. Tsegay, Effective pre-retrieval query performance prediction using similarity and variability evidence, in: C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, R. W. White (Eds.), Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings, volume 4956 of *Lecture Notes in Computer Science*, Springer, 2008, pp. 52–64. URL: https://doi.org/10.1007/978-3-540-78646-7_8. doi:10.1007/978-3-540-78646-7\\_8.

[12] Y. Zhou, W. B. Croft, Query performance prediction in web search environments, in: W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, N. Kando (Eds.), SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007, ACM, 2007, pp. 543–550. URL: https://doi.org/10.1145/1277741.1277835. doi:10.1145/1277741.1277835.

[13] A. Shtok, O. Kurland, D. Carmel, F. Raiber, G. Markovits, Predicting query performance by query-drift estimation, ACM Trans. Inf. Syst. 30 (2012) 11:1–11:35. URL: https://doi.org/10.1145/2180868.2180873. doi:10.1145/2180868.2180873.

[14] Y. Tao, S. Wu, Query performance prediction by considering score magnitude and variance together, in: J. Li, X. S. Wang, M. N. Garofalakis, I. Soboroff, T. Suel, M. Wang (Eds.), Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014, ACM, 2014, pp. 1891–1894. URL: https://doi.org/10.1145/2661829.2661906. doi:10.1145/2661829.2661906.

[15] S. Cronen-Townsend, Y. Zhou, W. B. Croft, Predicting query performance, in: K. Järvelin,

M. Beaulieu, R. A. Baeza-Yates, S. Myaeng (Eds.), SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland, ACM, 2002, pp. 299–306. URL: https://doi.org/10.1145/564376.564429. doi:10.1145/564376.564429.

[16] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, E. M. Voorhees, Overview of the TREC 2019 deep learning track, CoRR abs/2003.07820 (2020). URL: https://arxiv.org/abs/2003.07820. arXiv:2003.07820.

[17] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, Overview of the TREC 2020 deep learning track, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020, volume 1266 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2020. URL: https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.DL.pdf.

[18] M. Fröbe, A. Parry, H. Scells, S. Wang, S. Zhuang, G. Zuccon, M. Potthast, M. Hagen, Corpus Subsampling: Estimating the Effectiveness of Neural Retrieval Models on Large Corpora, in: Advances in Information Retrieval. 47th European Conference on IR Research (ECIR 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[19] A. Bondarenko, M. Fröbe, M. Beloucif, L. Gienapp, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché~2020: Argument retrieval, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Proceedings of CLEF 2020, volume 12260 of *LNCS*, Springer, 2020, pp. 384–395. doi:10.1007/978-3-030-58219-7_26.

[20] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: J. Vanschoren, S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2021. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.

[21] V. Boteva, D. G. Ghalandari, A. Sokolov, S. Riezler, A full-text learning to rank dataset for medical information retrieval, in: N. Ferro, F. Crestani, M. Moens, J. Mothe, F. Silvestri, G. M. D. Nunzio, C. Hauff, G. Silvello (Eds.), Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings, volume 9626 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 716–722. URL: https://doi.org/10.1007/978-3-319-30671-1_58. doi:10.1007/978-3-319-30671-1\\_58.

[22] C. Cleverdon, J. Mills, M. Keen, Factors Determining the Performance of Indexing Systems. Volume I. Design. Part 2. Appendices., Technical Report PB169574, Association of Special Libraries and Information Bureau, Cranfield (England)., 1966. URL: https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB169574.xhtml, num Pages: 261.

[23] M. E. Lesk, G. Salton, Relevance assessments and retrieval system evaluation, Inf. Storage Retr. 4 (1968) 343–359. URL: https://doi.org/10.1016/0020-0271(68)90029-6. doi:10.1016/0020-0271(68)90029-6.

[24] E. M. Voorhees, Variations in relevance judgments and the measurement of retrieval effectiveness, in: W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, J. Zobel (Eds.), SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia, ACM, 1998, pp. 315–323. URL: https://doi.org/10.1145/290941.291017. doi:10.1145/290941.291017.

[25] C. Meng, N. Arabzadeh, A. Askari, M. Aliannejadi, M. de Rijke, Query performance prediction using relevance judgments generated by large language models, CoRR abs/2404.01012 (2024). URL: https://doi.org/10.48550/arXiv.2404.01012. doi:10.48550/ARXIV.2404.01012. arXiv:2404.01012.