Limites et Robustesse de la Prédiction de Performance des Requêtes - Résumé étendu

Adrian-Gabriel Chifu[1] Sébastien Déjean[2] Moncef Garouani[3] Josiane Mothe[3]
Diégo Ortiz[3] Md Zia Ullah[4]
(1) Aix Marseille Université, Université de Toulon, CNRS, LIS, Marseille, France
(2) IMT, UMR5219 CNRS, UPS, Univ. de Toulouse, Toulouse, France
(3) IRIT, UMR 5505 CNRS, Univ. de Toulouse, Toulouse, France
(4) Edinburgh Napier University, Edinburgh, UK
[1]adrian.chifu@univ-amu.fr, [2]sebastien.dejean@math.univ-toulouse.fr,
[3]{moncef.garouani,josiane.mothe,diego.ortiz}@irit.fr,
[4]m.ullah@napier.ac.uk

RÉSUMÉ

La prédiction de performance des requêtes (QPP) est cruciale pour optimiser les systèmes de recherche d'information (RI). Ce papier présente une évaluation systématique de la robustesse des méthodes de QPP à travers différents paradigmes de modèles de recherche d'information (creux, denses, hybrides) et sur différentes collections. Nos résultats démontrent que la fiabilité des prédicteurs actuels est fortement dépendante de la collection et du moteur de recherche utilisé, limitant leur application pratique au traitement sélectif des requêtes.

ABSTRACT

Uncovering the Limitations of Query Performance Prediction : Failures, Insights, and Implications for Selective Query Processing

Query performance prediction (QPP) is crucial for optimising information retrieval systems. This paper presents a systematic evaluation of the robustness of QPP models across different information retrieval paradigms (sparse, dense, hybrid) and on different information benchmark collections. Our results demonstrate that the reliability of current predictors is highly dependent on the collection and retrieval model used, limiting their practical application to selective query processing.

MOTS-CLÉS : Recherche d'information, Prédiction de la performance des requêtes, Analyse complète.

KEYWORDS: Information retrieval, Query performance prediction, Comprehensive analysis.

Lorsque l'article fait l'objet d'une double soumission et que la réponse d'acceptation est déjà connue, l'article doit être déclarée "acceptée" selon le modèle ci-dessous en précisant le nom de la conférence ou de la revue ainsi que l'url de l'article si celui-ci est connu. Lorsque l'acceptation n'est pas connue, l'article doit être déclarée "soumis" en précisant seulement le nom de la conférence ou de la revue. Ne laisser aucune indication en cas de soumission de travaux originaux.

ARTICLE ACCEPTÉ À : ACM Transactions on Information Systems (TOIS) 2025 (Chifu *et al.*, 2025).
URL : https://dl.acm.org/doi/10.1145/3774427

1 Introduction

La prédiction de la performance des requêtes (QPP) vise à estimer l'efficacité d'un système de recherche d'information pour une requête donnée, sans disposer de jugements de pertinence humains (Carmel & Yom-Tov, 2010). Cette estimation joue un rôle crucial dans l'optimisation des moteurs de recherche, notamment pour le traitement sélectif des requêtes. En identifiant les requêtes difficiles, le système peut décider dynamiquement d'activer des mécanismes de correction, tels que l'expansion de requête (Amati *et al.*, 2004; Cronen-Townsend *et al.*, 2004) ou la sélection automatique du modèle de classement le plus approprié (Deveaud *et al.*, 2018).

Malgré des décennies de recherche, la robustesse et la capacité de généralisation des méthodes QPP restent mal comprises (Cronen-Townsend *et al.*, 2002; He & Ounis, 2004; Mothe & Tanguy, 2005; Hauff *et al.*, 2008; Carmel & Yom-Tov, 2010; Pérez-Iglesias & Araujo, 2010; Katz *et al.*, 2014; Thomas *et al.*, 2017; Déjean *et al.*, 2020; Chifu *et al.*, 2024; Meng *et al.*, 2024; Arabzadeh *et al.*, 2024; Saha *et al.*, 2025; Meng *et al.*, 2025). La littérature existante s'est principalement concentrée sur l'évaluation de prédicteurs (comme NQC ou WIG) dans le cadre de systèmes de recherche « creux » (sparse) traditionnels, tels que BM25 (Robertson, Stephen E *et al.*, 1995) ou les modèles à divergence DFree (Amati & Van Rijsbergen, 2002). Cependant, le développement des modèles de recherche neuronaux a radicalement transformé le paysage. L'apparition de modèles denses comme ColBERT (Khattab & Zaharia, 2020) ou de modèles hybrides comme SPLADE (Formal *et al.*, 2021) soulève une question fondamentale : les prédicteurs de performance actuels sont-ils fiables lorsqu'ils sont transposés d'un paradigme de recherche à un autre ?

Ce travail propose la première évaluation exhaustive et multi-paradigme de la robustesse de la QPP. Nous analysons la capacité de généralisation de divers prédicteurs, incluant des méthodes post-recherche classiques comme NQC (Shtok *et al.*, 2012), des approches basées sur l'apprentissage (LETOR) (Qin *et al.*, 2010; Chifu *et al.*, 2018), ainsi que des prédicteurs neuronaux récents (MQPPF, BERT-QPP). Notre étude couvre quatre collections de référence majeures (TREC Robust, GOV2, WT10G et MS-MARCO) et trois familles de modèles de classement : creux (BM25, DFree), denses (ColBERT, TCT-ColBERT (Lin *et al.*, 2020)) et hybrides (SPLADE).

Nos résultats mettent en évidence les limites des prédicteurs de performance. L'analyse révèle une forte variabilité de la précision des prédictions, la collection de référence utilisée étant le facteur qui influence le plus les résultats, suivi par le type de modèle de classement. En revanche, les résultats ne sont pas stables : les prédicteurs performants sur des collections échouent souvent sur des corpus Web complexes ou lorsqu'ils sont appliqués à des modèles denses. De plus, nous démontrons que dans un scénario d'application réelle — l'expansion sélective de requêtes — les méthodes s'appuyant sur les prédicteurs de la littérature ne permettent que des gains marginaux (environ 4 % d'amélioration du NDCG), soulignant le besoin de développer des prédicteurs plus universels et robustes aux changements d'architecture.

2 Protocole expérimental

Nous avons mis en place un protocole expérimental destiné à tester la robustesse et la capacité de généralisation des méthodes de QPP sur des contextes très variés. Les expériences s'appuient sur une ensemble de collections de référence couvrant des domaines variés, en utilisant les titres de sujets (*topics*) comme requêtes. Nous utilisons TREC Robust04 ($\approx$ 0,5 million de documents) et

WT10G (≈ 1,7 million). Le protocole s'étend ensuite aux corpus Web à grande échelle avec GOV2 (25 millions de documents) et MS-MARCO. En plus de la version *Passage* initiale (8,8 millions), nous intégrons les jeux de données des éditions TREC Deep Learning 2021 et 2022 (basés sur MS-MARCO v2), totalisant un volume de 138 millions de passages. Ce saut d'échelle est crucial pour évaluer la robustesse des prédicteurs face à des labels extrêmement épars. Côté moteurs de recherche, l'étude compare trois paradigmes d'ordonnancement : des méthodes basées sur des représentations non denses (lexicales comme BM25), hybrides (comme SPLADE) et denses (comme ColBERT), permettant d'analyser la transférabilité de la QPP entre architectures.

Pour les méthodes avec des représentations non denses, nous utilisons BM25 (Robertson, Stephen E *et al.*, 1995) et DFree (Amati *et al.*, 2004), implémentés dans Terrier (Macdonald *et al.*, 2013) avec les paramètres par défaut recommandés (Ounis *et al.*, 2005), et évalués aussi avec expansion automatique de requête via Bo2 (Amati & Van Rijsbergen, 2002). Pour les approches neuronales, nous considérons SPLADE v2 / distilSPLADE (Formal *et al.*, 2021, 2022), ainsi que ColBERT et sa version ColBERTv2 (Khattab & Zaharia, 2020; Santhanam *et al.*, 2022), et TCT-ColBERT / TCT-ColBERT-v2 (Lin *et al.*, 2020, 2021). Les modèles denses sont évalués via le cadre BEIR (Thakur *et al.*, 2021), et certaines variantes sont aussi fine-tunées pour le réordonnancement.

Pour la prédiction, nous avons implémenté et comparé trois familles distinctes de signaux, totalisant une couverture exhaustive des approches actuelles :

1. *Prédicteurs post-recherche classiques (SOTA) :* Nous utilisons les standards du domaine tels que NQC (Shtok *et al.*, 2009) (variabilité normalisée des scores), UQC (Shtok *et al.*, 2012), WIG (gain d'information pondéré) et QF (Zhou & Croft, 2007). Ces méthodes reposent sur l'hypothèse que la distribution des scores de pertinence (dispersion, asymétrie) révèle la confiance du moteur.
2. *Caractéristiques LETOR agrégées :* Inspirés par nos travaux précédents (Chifu *et al.*, 2018), nous exploitons les scores bruts de correspondance entre les requêtes et les documents issus des modèles de pondération. Contrairement à l'approche learning-to-rank classique, nous agrégeons ces scores (moyenne, maximum, variance) sur les k premiers documents pour les utiliser comme des prédicteurs au niveau des requêtes. Ces caractéristiques sont ensuite normalisées par la longueur effective de la requête (nombre de termes présents dans l'index) pour éviter les biais de longueur.
3. *Prédicteurs neuronaux denses :* Nous évaluons BERT-QPP (Arabzadeh *et al.*, 2021) dans ses deux variantes : le *Bi-Encoder* (B_{bi}), plus rapide, qui encode la requête et les documents séparément, et le *Cross-Encoder* (B_{cross}), plus précis mais coûteux, qui capture les interactions fines. Nous incluons également le modèle multi-tâches MQPPF (Khodabakhsh & Bagheri, 2023), qui apprend conjointement à classer les documents et à prédire la performance.

L'évaluation repose principalement sur des corrélations entre prédictions et efficacité (NDCG, MAP), afin de comparer finement collections, modèles de recherche et métriques.

3 Résultats principaux

Les principaux résultats montrent une fragilité structurelle des approches de QPP dès que l'on change de contexte (collection, paradigme d'ordonnancement, métrique). D'abord, même dans leur cadre « naturel » (ordonnancement basé sur des méthodes non denses de type BM25 / DFree), les prédicteurs

classiques comme NQC/UQC, WIG, QF n'atteignent le plus souvent que des corrélations modestes avec l'efficacité, et ces corrélations varient fortement selon les collections.

Ensuite, le passage à des moteurs hybrides ou denses (SPLADE, ColBERT, TCT-ColBERT) dégrade nettement les prédicteurs conçus pour les modèles non dense, et les prédicteurs « denses » (BERT-QPP, MQPPF) ne résolvent pas le problème : ils peuvent être utiles dans certains cas, mais ne généralisent pas de manière fiable entre paradigmes (dense vs. non dense).

Pour comprendre ces échecs de généralisation, il est nécessaire d'analyser le comportement des prédicteurs isolés en fonction de la nature des collections. Nous distinguons ici les corpus journalistiques traditionnels, représentés par TREC Robust04, des collections de passages Web massives et bruitées comme MS-MARCO. La table 1 (issue de la table 4 de l'article original) présente une analyse détaillée des corrélations de Pearson (r) et de Kendall (τ) pour chaque famille de prédicteurs.

Nous observons une dichotomie marquée. D'une part, les caractéristiques LETOR qui correspondent à l'agrégation des scores de ressemblance entre documents et requêtes permettent de bien prédire les performance du modèle BM25 sur TREC Robust04 ($r = 0.459$ pour L.BM25). D'autre part, ces mêmes indicateurs ne sont pas efficaces lorsqu'ils sont appliqués au contexte Web de MS-MARCO ($r = 0.149$).

De plus, le tableau met en évidence une « spécialisation architecturale » des prédicteurs neuronaux. Le prédicteur B_{bi} (BERT bi-encoder) est très efficace pour estimer la qualité d'un classement dense produit par TCT-ColBERT ($r = 0.507$), mais devient inopérant ($r = 0.151$) pour prédire la qualité d'un classement lexical (BM25). Cela confirme que le signal de difficulté pourrait être intrinsèquement lié à la nature (lexicale ou sémantique) du modèle de recherche ciblé.

NDCG	**BM25**				**SPLADE**				**TCT-ColBERT**				**ColBERT**			
Predictor	Robust		MS-MC.		Robust		MS-MC.		Robust		MS-MC.		Robust		MS-MC.	
	r	τ	r	τ	r	τ	r	τ	r	τ	r	τ	r	τ	r	τ
L.BM25	.459‡	.321‡	.149	.120	.279‡	.209‡	.162	.139†	.207†	.101†	.178	.107	.157†	.121†	-.034	-.152†
L.DFree	.443‡	.290‡	.157	.116	.268‡	.218‡	.182	.131	.187†	.070	.193	.110	.144†	.135†	.014	-.128
L.Lemur	.456‡	.326‡	.121	.103	.200†	.163‡	.125	.118	.122	.112†	.102	.062	.062	.074	-.068	-.199†
L.InExp2	.424‡	.327‡	.070	.100	.283‡	.228‡	.187	.157†	.231‡	.102†	.211†	.143†	.208†	.158‡	.010	-.144†
UQC	.407‡	.322‡	-.123	-.025	.439‡	.328‡	.401‡	.277‡	.468‡	.336‡	-.068	.058	.488‡	.329‡	-.394‡	-.494‡
NQC	.354‡	.285‡	-.010	-.005	.295‡	.226‡	.212†	.156†	.265‡	.181‡	-.075	-.006	.254‡	.177‡	-.384‡	-.601‡
WIG	.342‡	.236‡	.027	-.080	.354‡	.236‡	.179	.086	.436‡	.287‡	.208†	.087	.477‡	.315‡	.848‡	.692‡
QF	.394‡	.265‡	.146	.106	.436‡	.297‡	.418‡	.320‡	.447‡	.301 ‡	.213 †	.102	.456‡	.318 ‡	.811‡	.543‡
B_{bi}	.151†	.104†	-.166†	.157‡	.177†	.082	.204†	-.078	.507‡	.337‡	.186	.091	.495‡	.334‡	.608‡	.437‡
B_{cross}	.069	.034	.009	-.042	.048	.019	.006	.111	.044	.021	-.041	.037	-.050	-.030‡	-.083	-.010
MQPPF	.237‡	.163‡	.436‡	.316‡	.225‡	.137‡	.230†	.150†	.184‡	.125‡	.280‡	.210‡	.202‡	.121‡	.713‡	.304‡

TABLE 1 – Corrélations individuelles (Pearson r et Kendall τ) entre NDCG et les prédictions sur TREC Robust et MS-MARCO (adapté de la table 4 de l'article original). † et ‡ : $p < 0.05$ et < 0.01.

Outre la variabilité due aux collections, nos résultats montrent que les corrélations sont systématiquement plus élevées avec des métriques globales (NDCG/MAP) qu'avec des métriques de précision pure (P@10), suggérant un meilleur alignement des signaux actuels avec une notion de gain cumulé.

Face aux limites des prédicteurs individuels, une question centrale de notre étude est de savoir si la combinaison supervisée de plusieurs signaux permet d'obtenir une prédiction plus robuste. Nous avons pour cela évalué deux modèles d'apprentissage : la régression linéaire (LR), qui capture les relations simples, et les forêts aléatoires (RF), capables de modéliser des dépendances non linéaires complexes. Nous avons appliqué ces modèles à trois ensembles de caractéristiques : le groupe «

SOTA », le groupe « LETOR » et l'ensemble « BOTH », qui réunit la totalité des signaux disponibles. L'objectif est de vérifier si l'enrichissement de l'espace de caractéristiques (BOTH) couplé à un modèle non linéaire (RF) permet de mieux prédire des métriques comme la précision en haut de liste (P@10).

La table 2 (table 7 de l'article original) présente les résultats de cette analyse. Elle confirme que si l'approche « BOTH + RF » améliore la prédiction du NDCG ($r = 0.492$ sur BM25), elle échoue à corriger le biais structurel défavorable à la P@10. La corrélation avec P@10 reste systématiquement inférieure, voire inexistante pour les modèles denses comme SPLADE ($r = 0.054$ avec LETOR).

Au-delà des corrélations, deux analyses complémentaires éclairent la nature du problème. D'une part, nous avons quantifié l'erreur absolue de prédiction via la MAE (*Mean Absolute Error*) et le RMSE (table 11 de l'article original). Les résultats sont préoccupants : même lorsque la corrélation est modérée ($r \approx 0.4$), l'erreur de prédiction reste souvent élevée, indiquant que le prédicteur parvient à classer les requêtes (laquelle est plus difficile que l'autre) mais échoue à estimer le score réel de précision.

D'autre part, pour isoler les causes de cette instabilité, nous avons mené une analyse de variance (ANOVA) détaillée (table 10 de l'article original). Cette analyse statistique révèle que le facteur « Collection » explique une part prépondérante de la variance totale des performances, loin devant le facteur « Ranker » (système de recherche) ou l'interaction entre les deux. En d'autres termes, la difficulté intrinsèque des requêtes et des documents dictent la performance du QPP bien plus que l'architecture du moteur de recherche lui-même. Cela explique pourquoi un prédicteur robuste sur TREC Robust peut s'effondrer sur MS-MARCO, indépendamment du système de recherche.

r	Mod.	**BM25**			**SPLADE**			**ColBERT**			**TCT-ColBERT**		
		NDCG	MAP	P@10	NDCG	MAP	P@10	NDCG	MAP	P@10	NDCG	MAP	P@10
SOTA	LR	.446‡	.322‡	.211‡	.511‡	.433‡	.355‡	.483‡	.407‡	.388‡	.473‡	.342‡	.291‡
-	RF	.432‡	.287‡	.243‡	.447‡	.369‡	.318‡	.470‡	.369‡	.324‡	.438‡	.278‡	.244‡
LETOR	LR	.424‡	.489‡	.326‡	.303‡	.266‡	.054	.209‡	.228‡	.126†	.184‡	.117	.098
-	RF	.400‡	.382‡	.298‡	.224‡	.208‡	.124†	.142†	.059	-.009	.206‡	.088	-.043
BOTH	LR	.449‡	.478‡	.273‡	.488‡	.421‡	.313‡	.433‡	.391‡	.331‡	.447‡	.327‡	.255‡
-	RF	.492‡	.441‡	.356‡	.491†	.393‡	.315‡	.499‡	.357‡	.303‡	.438‡	.263‡	.206‡

TABLE 2 – Corrélation combinée (r) sur TREC Robust pour NDCG, MAP et P@10 (adapté de la table 7 de l'article original).

Ces faiblesses dans les prédictions ne sont pas de simples artefacts statistiques ; elles se traduisent directement par une inefficacité dans les applications en aval. Pour évaluer l'utilité réelle, nous avons simulé un scénario de traitement sélectif des requêtes (*Selective Query Processing* (Mothe & Ullah, 2023)). La tâche consiste à utiliser le prédicteur pour décider dynamiquement, pour chaque requête, s'il est préférable d'utiliser le système de base (ex : BM25) ou une version plus coûteuse (ex : BM25 avec expansion de requête ou un modèle neuronal).

Les résultats de cette expérience (détaillés dans la Figure 4 de l'article original) sont révélateurs. Alors qu'un choix optimal (« Oracle ») permettrait un gain théorique substantiel sur la mesure NDCG, la sélection guidée par les QPP n'apporte que des gains marginaux ($\approx 4\%$ d'amélioration sur TREC Robust). Par ailleurs, sur les collections difficiles comme MS-MARCO, le risque de dégradation de la performance est réel : les prédicteurs échouent souvent à identifier la configuration optimale, conduisant le système à choisir un modèle moins performant pour une requête donnée. Cela confirme que la fiabilité opérationnelle des QPP reste un défi majeur, la corrélation ne garantissant pas la

capacité de décision binaire.

Au-delà des analyses ciblées présentées ici, l'article original (Chifu *et al.*, 2025) développe un cadre expérimental plus vaste, nécessaire pour appréhender la complexité du problème. Il fournit d'abord les éléments de contexte indispensables : la table 1 détaille les propriétés statistiques des collections (taille, longueur des requêtes, densité des jugements de pertinence), la table 2 présente les architectures des systèmes de recherche évalués (non denses, hybrides, denses), tandis que la table 3 établit les niveaux de performance de référence (« baselines ») que les QPP tentent de capturer. La figure 1 illustre la variabilité des relations prédiction $\leftrightarrow$ performance, mettant en lumière les échecs de généralisation selon les configurations.

Dans l'article original, l'analyse approfondie se poursuit à travers plusieurs axes complémentaires. La figure 2 de l'article original propose une matrice de corrélations globale qui synthétise les dépendances structurelles entre familles de prédicteurs et mesures d'efficacité. Concernant les évaluations quantitatives, les tables 5 à 8 élargissent l'étude à d'autres métriques et combinaisons de caractéristiques. La robustesse est spécifiquement rapportée dans la table 9, qui analyse la sensibilité aux variations fines des systèmes (BM25 vs DFree, impact de l'expansion), tandis que l'ANOVA (table 10) et la figure 3 quantifient statistiquement la prédominance du facteur « collection » sur le facteur « moteur ». Enfin, pour évaluer l'utilité réelle en aval, la figure 4 simule des scénarios de sélection automatique de système, et la table 11 complète les corrélations par une analyse d'erreur (MAE/RMSE/R^2), confirmant que la fiabilité opérationnelle reste un défi majeur malgré certaines corrélations positives.

4 Conclusion

En conclusion, ce travail montre la difficulté de prédire les performances des requêtes hors d'un cadre spécifique : les prédicteurs. En effet, les prédicteurs, qu'ils soient « classiques » ou neuronaux, montrent des résultats instables et ne se généralisent pas entre collections, moteurs de recherche et métriques. Nos analyses confirment que la collection est le premier facteur de la variabilité observée, ce qui réduit la portée des comparaisons toutes choses égales par ailleurs et complique l'usage opérationnel des prédicteurs de performance pour piloter des décisions (expansion automatique de requêtes, sélection automatique de la configuration du moteur, etc.).

Compte tenu de ces résultats, nous pouvons formuler plusieurs recommandations à la communauté. Premièrement, l'évaluation de nouvelles méthodes de QPP ne devrait pas se limiter à une ou deux collections, mais au contraire couvrir plusieurs types d'ensembles de données caractérisés par des domaines et des tailles distincts, afin de vérifier la généralisation des prédicteurs à différents contextes. Deuxièmement, il est essentiel de ne pas se fier exclusivement aux corrélations globales et à leurs p-valeurs : une analyse détaillée des résultats révèle que des corrélations statistiquement significatives peuvent masquer une relation très bruitée ou dispersée. Il convient donc de compléter l'analyse des corrélations par des visualisations (nuages de points) et des métriques d'erreur (MAE, RMSE, MedAE, R^2), afin d'évaluer de manière réaliste la qualité prédictive. Troisièmement, lorsque des gains sur des tâches en aval sont revendiquées (par exemple pour la sélection de moteur ou l'expansion sélective), il est crucial de quantifier clairement les améliorations obtenues par rapport aux systèmes de base : dans cette étude, même dans les contextes les plus favorables, les gains ne dépassent qu'exceptionnellement 4 % en NDCG, ce qui pose la question de la valeur ajoutée réelle de QPP pour les déploiements opérationnels. Enfin, nos résultats mettent en évidence que la collection est le facteur prédominant

affectant la performance des prédicteurs (ANOVA : F = 137,6, p-valeur $< 2e{-}16$), bien devant le type de moteur (F = 3,0, p-valeur = 0,01). Cela implique qu'un recalibrage ou un réapprentissage est indispensable pour s'adapter à une nouvelle collection ou à un nouveau moteur de classement.

La conclusion principale est donc double : (i) il est difficile de concevoir des prédicteurs intrinsèquement robustes au changement de domaine et adaptés aux architectures denses et hybrides, et (ii) il est nécessaire d'évaluer ces méthodes dans des cadres dépassant les seules corrélations (erreurs, analyses distributionnelles), afin de relier plus directement la qualité de prédiction à l'utilité en aval.

Références

AMATI G., CARPINETO C. & ROMANO G. (2004). Query difficulty, robustness, and selective application of query expansion. In S. MCDONALD & J. TAIT, Éds., *26th European Conference on IR Research, (ECIR)*, p. 127–137, New York, USA : Springer Berlin Heidelberg.

AMATI G. & VAN RIJSBERGEN C. J. (2002). Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *Transactions on Information Systems (TOIS)*, **20**(4), 357–389.

ARABZADEH N., KHODABAKHSH M. & BAGHERI E. (2021). BERT-QPP : Contextualized Pre-Trained Transformers for Query Performance Prediction. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, p. 2857–2861, New York, USA : ACM.

ARABZADEH N., MENG C., ALIANNEJADI M. & BAGHERI E. (2024). Query performance prediction : From fundamentals to advanced techniques. In *European Conference on Information Retrieval*, p. 381–388, New York, USA : Springer.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

CARMEL D. & YOM-TOV E. (2010). Estimating the query difficulty for information retrieval.

CHIFU A.-G., DEJEAN S., GAROUANI M., MOTHE J., ORTIZ D. & ULLAH M. Z. (2024). Can we predict qpp ? an approach based on multivariate outliers. In *European Conference on Information Retrieval*, p. 458–467, New York, USA : Springer.

CHIFU A.-G., DÉJEAN S., GAROUANI M., MOTHE J., ORTIZ D. & ULLAH M. Z. (2025). Uncovering the limitations of query performance prediction : Failures, insights, and implications for selective query processing. *ACM Transactions on Information Systems*. DOI : https ://doi.org/10.1145/3774427.

CHIFU A.-G., LAPORTE L., MOTHE J. & ULLAH M. Z. (2018). Query performance prediction focused on summarized letor features. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 1177–1180, New York, USA : ACM.

CRONEN-TOWNSEND S., ZHOU Y. & CROFT W. B. (2002). Predicting query performance. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 299–306, New York, USA : ACM.

CRONEN-TOWNSEND S., ZHOU Y. & CROFT W. B. (2004). *A Language Modeling Framework for Selective Query Expansion*. Rapport interne, Technical Report IR-338, Center for Intelligent Information Retrieval ….

DÉJEAN S., IONESCU R. T., MOTHE J. & ULLAH M. Z. (2020). Forward and backward feature selection for query performance prediction. In *Proceedings of the 35th Annual ACM symposium on applied computing*, p. 690–697, New York, USA : ACM.

DEVEAUD R., MOTHE J., ULLAH M. Z. & NIE J.-Y. (2018). Learning to adaptively rank document retrieval system configurations. *ACM Transactions on Information Systems (TOIS)*, **37**(1), 3. DOI : 10.1145/3231937.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

FORMAL T., LASSANCE C., PIWOWARSKI B. & CLINCHANT S. (2021). Splade v2 : Sparse lexical and expansion model for information retrieval.

FORMAL T., LASSANCE C., PIWOWARSKI B. & CLINCHANT S. (2022). From distillation to hard negative sampling : Making sparse neural ir models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, p. 2353–2359, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3477495.3531857.

HAUFF C., HIEMSTRA D. & DE JONG F. (2008). A survey of pre-retrieval query performance predicors. In *Proceedings of the 17th ACM International Conference on Information and Knowledge Management*, p. 1419–1420, New York, USA : ACM.

HE B. & OUNIS I. (2004). Inferring query performance using pre-retrieval predicors. In *International Symposium on String Processing and Information Retrieval*, p. 43–54, New York, USA : Springer.

KATZ G., SHTOK A., KURLAND O., SHAPIRA B. & ROKACH L. (2014). Wikipedia-based query performance prediction. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 1235–1238, New York, USA : ACM.

KHATTAB O. & ZAHARIA M. (2020). Colbert : Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, p. 39–48, New York, USA : ACM. DOI : 10.1145/3397271.3401075.

KHODABAKHSH M. & BAGHERI E. (2023). Learning to rank and predict : Multi-task learning for adhoc retrieval and query performance prediction. *Information Sciences*, **639**, 119015.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LIN S.-C., YANG J.-H. & LIN J. (2020). Distilling dense representations for ranking using tightly-coupled teachers.

LIN S.-C., YANG J.-H. & LIN J. (2021). In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, p. 163–173, New York, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2021.repl4nlp-1.17.

MACDONALD C., SANTOS R. L., OUNIS I. & HE B. (2013). About learning models with multiple query-dependent features. *Transactions on Information Systems (TOIS)*, **31**(3), 11.

MENG C., ARABZADEH N., ASKARI A., ALIANNEJADI M. & DE RIJKE M. (2024). Query performance predicion using relevance judgments generated by large language models.

MENG C., FAGGIOLI G., ALIANNEJADI M., FERRO N. & MOTHE J. (2025). Qpp++ 2025 : Query performance prediction and its applications in the era of large language models. In *European Conference on Information Retrieval*, p. 319–325 : Springer.

MOTHE J. & TANGUY L. (2005). Linguistic Features to Predic Query Difficulty. In *Predicting Query Difficulty, SIGIR workshop*, p. 7–10, New York, USA : ACM.

MOTHE J. & ULLAH M. Z. (2023). Selective query processing : A risk-sensitive selection of search configurations. *ACM Transactions on Information Systems*, **42**(1), 1–35.

OUNIS I., AMATI G., PLACHOURAS V., HE B., MACDONALD C. & JOHNSON D. (2005). Terrier information retrieval platform. In *Advances in Information Retrieval : 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*, p. 517–519, New York, USA : Springer.

PÉREZ-IGLESIAS J. & ARAUJO L. (2010). Standard deviation as a query hardness estimator. In *International Symposium on String Processing and Information Retrieval*, p. 207–212, New York, USA : Springer.

QIN T., LIU T.-Y., XU J. & LI H. (2010). LETOR : A Benchmark Collection for Research on Learning to Rank for Information Retrieval. *Information Retrieval*, **13**(4), 346–374.

ROBERTSON, STEPHEN E, WALKER, STEVE, JONES, SUSAN, HANCOCK-BEAULIEU, MICHELINE M, GATFORD, MIKE & OTHERS (1995). *Okapi at TREC-3*. 183 Euston Road London NW1 2BE : British Library Research and Development Department.

SAHA S., DATTA S., ROY D., MITRA M. & GREENE D. (2025). Combining Query Performance Predictors : A Reproducibility Study. In *European Conference on Information Retrieval*, p. 112–129, New York, USA : Springer.

SANTHANAM K., KHATTAB O., SAAD-FALCON J., POTTS C. & ZAHARIA M. (2022). Colbertv2 : Effective and efficient retrieval via lightweight late interaction. DOI : 10.18653/v1/2022.naacl-main.272.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SHTOK A., KURLAND O. & CARMEL D. (2009). Predicting query performance by query-drift estimation. In *Advances in Information Retrieval Theory : Second International Conference on the Theory of Information Retrieval, ICTIR 2009 Cambridge, UK, September 10-12, 2009 Proceedings 2*, p. 305–312, New York, USA : Springer.

SHTOK A., KURLAND O., CARMEL D., RAIBER F. & MARKOVITS G. (2012). Predicting query performance by query-drift estimation. *Transactions on Information Systems (TOIS)*, **30**(2), 11.

THAKUR N., REIMERS N., RÜCKLÉ A., SRIVASTAVA A. & GUREVYCH I. (2021). BEIR : A heterogeneous benchmark for zero-shot evaluation of information retrieval models.

THOMAS P., SCHOLER F., BAILEY P. & MOFFAT A. (2017). Tasks, queries, and rankers in pre-retrieval performance predicion. In *Proceedings of the 22nd Australasian Document Computing Symposium*, p. 1–4, New York, USA : ACM.

ZHOU Y. & CROFT W. B. (2007). Query performance predicion in web search environments. In *ACM SIGIR*, p. 543–550, New York, USA : ACM.