Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

iQPP: Une Référence pour la Prédiction de Performances des Requêtes d'Images

Eduard Poesina[1] Radu Tudor Ionescu[1] Josiane Mothe[2]
(1) Department of Computer Science, University of Bucharest, 14 Academiei, Bucharest, Romania
(2) INSPE, IRIT UMR5505 CNRS, Université Toulouse Jean-Jaurès, 118 Rte de Narbonne, Toulouse, France
eduardgabriel.poe@gmail.com, raducu.ionescu@gmail.com,
josiane.mothe@irit.fr

RÉSUMÉ

La prédiction de la performance des requêtes (QPP) dans le contexte de la recherche d'images basée sur le contenu reste une tâche largement inexplorée, en particulier dans le scénario de la recherche par l'exemple, où la requête est une image. Pour stimuler les recherches dans ce domaine, nous proposons la première collection de référence. Nous proposons un ensemble de quatre jeux de données (PASCAL VOC 2012, Caltech-101, ROxford5k et RParis6k) avec les performances attendues pour chaque requête à l'aide de deux modèles de recherche d'images état de l'art. Nous proposons également de nouveaux prédicteurs pré et post-recherche. Les résultats empiriques montrent que la plupart des prédicteurs ne se généralisent pas aux différents scénarios d'évaluation. Nos expériences exhaustives indiquent que l'iQPP est une référence difficile, révélant une importante lacune dans la recherche qui doit être abordée dans les travaux futurs. Nous publions notre code et nos données [1]. Il s'agit du résumé étendu d'une publication acceptée à SIGIR 2023 (Poesina *et al.*, 2023).

ABSTRACT

iQPP : A Benchmark for Image Query Performance Prediction

Query performance prediction (QPP) in the context of content-based image retrieval remains a largely unexplored task, especially in the query-by-example scenario, where the query is an image. To stimulate research in this area, we propose the first benchmark. We propose a set of four datasets (PASCAL VOC 2012, Caltech-101, ROxford5k, and RParis6k) and estimate the ground-truth difficulty of each query using two state-of-the-art image retrieval models. We also propose new pre- and post-retrieval predictors. The empirical results show that most predictors do not generalize to different evaluation scenarios. Our extensive experiments indicate that iQPP is a challenging benchmark, revealing an important research gap that must be addressed in future work. We publish our code and data[1]. This is an extended abstract from a paper published at SIGIR 2023 (Poesina *et al.*, 2023).

MOTS-CLÉS : Systèmes d'information, Recherche d'information, prédiction de performance des requêtes, recherche d'images .

KEYWORDS: Information systems, Information retrieval, Query performance prediction, Content-based image retrieval.

1. https://github.com/Eduard6421/iQPP

1 Introduction

La prédiction des performances des requêtes (QPP) est la tâche qui consiste à estimer l'efficacité d'une recherche obtenue en réponse à une requête par un moteur de recherche, sans jugement de pertinence (Cronen-Townsend *et al.*, 2002). L'importance de cette tâche est reconnue en recherche d'information (Cronen-Townsend *et al.*, 2002; He & Ounis, 2004; Mothe & Tanguy, 2005; Hauff *et al.*, 2008, 2009; Shtok *et al.*, 2010; Cummins *et al.*, 2011; Kurland *et al.*, 2012; Cummins, 2014; Katz *et al.*, 2014; Raiber & Kurland, 2014; Roitman *et al.*, 2017; Chifu *et al.*, 2018; Mizzaro *et al.*, 2018; Roitman, 2018; Zamani *et al.*, 2018; Roy *et al.*, 2019; Arabzadeh *et al.*, 2020; Déjean *et al.*, 2020), et intéresse actuellement la communauté scientifique (Chen *et al.*, 2022; Datta *et al.*, 2022; Faggioli *et al.*, 2022; Jafarzadeh & Ensan, 2022). Cependant, dans le contexte de la recherche d'images, la prédiction de la performance des requêtes à retenu moins d'attention jusqu'ici, avec seulement quelques travaux publiés (Xing *et al.*, 2010; Li *et al.*, 2012; Nie *et al.*, 2012; Tian *et al.*, 2012; Jia *et al.*, 2014; Jia & Tian, 2015; Tian *et al.*, 2015; Pedronette & Torres, 2015; Sun *et al.*, 2018; Valem & Pedronette, 2021). Très peu de papiers ont par ailleurs considéré l'angle des requêtes par l'exemple, scénario dans lequel la requête est une image (Li *et al.*, 2012; Pedronette & Torres, 2015; Sun *et al.*, 2018; Valem & Pedronette, 2021).

Nous considérons que l'étude de la prédiction de performance est tout aussi importante pour l'image que pour le texte. Afin de développer l'intérêt de la communauté scientifique pour contexte de la recherche d'images basée sur le contenu, où les images doivent être retrouvées à partir d'une requête image, nous avons développé une collection de référence complète que nous appelons iQPP. Elle comprend quatre ensembles de données (PASCAL VOC 2012 (Everingham *et al.*, 2015), Caltech-101 (Li *et al.*, 2022), ROxford5k (Radenović *et al.*, 2018) et RParis6k (Radenović *et al.*, 2018)), deux systèmes de recherche d'images (Radenović *et al.*, 2019; Revaud *et al.*, 2019), ainsi que plusieurs prédicteurs de performance de requête pré- et post- recherche, pour lesquels nous fournissons les niveaux de performance prédits et réels pour deux mesures d'efficacité.

Les sections suivantes résume les ressources constituant la référence iQPP. Un descriptif plus complet se trouve dans la publication orginale de Poesina *et al.* (2023) ainsi que sur le GitHub `https://github.com/Eduard6421/iQPP`.

2 Ensembles d'images

Les quatre ensembles d'images sont PASCAL VOC 2012 (Everingham *et al.*, 2015), Caltech-101 (Li *et al.*, 2022), ROxford5k (Radenović *et al.*, 2018) et RParis6k (Radenović *et al.*, 2018). ROxford5k et RParis6k sont reconnus dans le cadre de la recherche d'images ; ils comprennnet chacun 70 requêtes. Nous avons par ailleurs adapté PASCAL VOC 2012 et Caltech-101 à la tâche de prédiction de performance. Nous avons créé 700 requêtes d'apprentissage et 700 de test pour chacun (Voir Table 1 et Figure 1).

3 Modèle de recherche et évaluation

Pour évaluer la difficulté d'une requête, nous avons considéré deux mesures d'efficacité de la recherche, la précision moyenne (AP) et la précision pour les k premiers résultats retrouvés (P@k).

TABLE 1 – Informations sur les jeux de données de la référence iQPP : nombre d'images, de requêtes d'entrainement et test pour chacun.

Jeux de données	#images	#requêtes d'entrainement	#requêtes de test
PASCAL VOC 2012	17,125	700	700
Caltech-101	9,146	700	700
ROxford5k	5,063	-	70
RParis6k	6,392	-	70

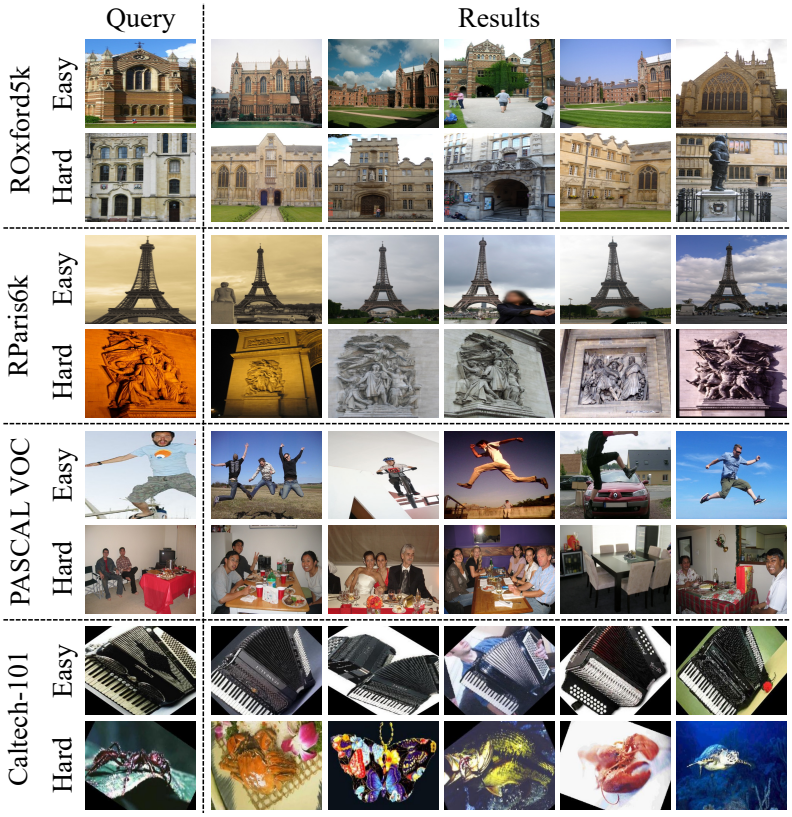

FIGURE 1 – Exemples de requêtes faciles (performance élevée) et de requêtes difficiles (faible performance) issues des 4 jeux d'images de notre référence iQPP. Pour chaque requête, nous montrons les cinq premiers résultats renvoyés par le système de Radenović et al. (Radenović *et al.*, 2019) pour illustrer les niveaux de performance des requêtes choisies. Cette image issue de (Poesina *et al.*, 2023) gagne a être vue en couleur.

Bien que la précision P@10 soit généralement utilisée dans le cadre de la prédiction de la performance des requêtes textuelles, nous avons constaté qu'un pourcentage élevé de requêtes de test dans les collections d'images (entre 29% et 82%) ont un score P@10 de 1. Pour une meilleure estimation de la difficulté de la requête, nous avons décidé d'utiliser plutôt la précision P@100.

Le premier modèle de recherche d'images que nous utilisons a été proposé par Radenović et al. (Radenović *et al.*, 2019) [2]. Il s'agit d'un modèle à base de réseau neuronal convolutif. Le second modèle

2. https://github.com/filipradenovic/cnnimageretrieval-pytorch

a été présenté par Revaud et al. (Revaud *et al.*, 2019) [3]. Le système s'appuie sur ResNet-101 (He *et al.*, 2016) pré-entraîné sur ImageNet (Russakovsky *et al.*, 2015).

Pour estimer l'efficacité d'un prédicteur, nous utilisons les coefficients de corrélation de Pearson et de Kendall τ entre les niveaux d'efficacité prédits et réels des requêtes de test, en suivant la procédure d'évaluation usuelle pour cette tâche (Yom-Tov *et al.*, 2005; Zhao *et al.*, 2008; Chifu *et al.*, 2018; Faggioli *et al.*, 2022). Nous avons utilisé un test de Student à un niveau de confiance de $0,01$ (Roitman, 2018). La figure 2 présente les résultats obtenus sur deux collections. Les prédicteurs considérés sont décrits dans la publication originale, certains sont issus de l'état de l'art (Ionescu *et al.*, 2016; Soviany *et al.*, 2021; Sun *et al.*, 2018), d'autres sont propres à cette recherche.

Type	Supervised	Method	PASCAL VOC 2012								Caltech-101							
			Radenović et al. [50]				Revaud et al. [53]				Radenović et al. [50]				Revaud et al. [53]			
			AP		P@100		AP		P@100		AP		P@100		AP		P@100	
			Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
		Random	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pre-retrieval		#objects / area [67]	0.02	**0.22**	0.03	**0.25**	0.02	**0.27**	0.03	**0.25**	0.01	0.08	0.01	0.04	0.04	0.06	0.03	0.04
		Image difficulty [30]	**0.25**	0.19	**0.33**	0.23	**0.32**	0.24	**0.31**	0.22	−0.01	−0.02	−0.07	−0.07	0.00	−0.02	−0.07	−0.06
		Denoising AE	0.15	0.16	0.06	0.08	0.11	0.12	0.08	0.09	0.03	0.02	0.06	0.03	0.12	0.07	0.13	0.07
		Masked AE	0.11	0.11	0.01	0.05	0.01	0.06	−0.01	0.03	−0.04	−0.04	0.01	0.00	0.03	0.02	0.09	0.05
		Class head kurtosis	0.05	0.08	0.09	0.07	0.12	0.09	0.12	0.08	0.16	0.17	0.26	0.30	0.23	0.17	0.13	0.10
		Class head dispersion	0.08	0.09	0.13	0.08	0.17	0.11	0.17	0.10	0.25	0.20	**0.48**	**0.38**	0.32	0.23	0.21	0.15
		Cluster density	0.13	0.12	0.00	0.01	−0.02	−0.04	−0.01	−0.01	0.15	0.09	0.41	0.24	−0.13	0.09	−0.03	−0.4
	✓	Fine-tuned ViT	0.04	0.02	0.20	0.10	0.17	0.06	0.14	0.05	**0.54**	**0.38**	0.27	0.15	**0.65**	**0.47**	**0.41**	**0.20**
Post-retrieval		Score Variance [13]	0.02	0.05	−0.02	0.02	0.23	0.19	0.26	0.20	0.11	0.01	0.21	0.01	0.51	0.51	0.30	0.39
	✓	Correlation CNN [68]	0.27	0.07	0.32	0.16	0.32	0.15	0.26	0.11	**0.83**	**0.65**	**0.76**	**0.51**	**0.78**	**0.60**	**0.71**	**0.50**
		Adapted query feedback	0.23	0.16	0.37	0.21	0.41	0.26	0.41	0.24	0.60	0.43	0.60	0.46	0.56	0.40	0.60	0.44
		Iterative removal	0.16	0.13	0.35	0.20	0.41	0.26	0.40	0.23	0.57	0.41	0.57	0.42	0.31	0.20	0.40	0.23
		Embedding Variance	0.29	0.20	0.33	0.21	0.43	0.22	0.37	0.20	0.28	0.20	0.49	0.28	0.26	0.18	0.49	0.26
	✓	Meta-regressor	**0.36**	**0.28**	**0.45**	**0.29**	**0.51**	**0.34**	**0.48**	**0.30**	0.71	0.53	0.72	**0.51**	0.76	0.57	0.70	0.49

FIGURE 2 – Corrélation entre les prédicteurs constituant la référence et les performances constatées des modèles de recherche - Issu de (Poesina *et al.*, 2023)

4 Conclusion

Dans cet article qui est un résumé étendu de la publication de Poesina *et al.* (2023) à la conférence SIGIR 2023, nous avons présenté la première collection pour la prédiction de performance de requêtes dans le cadre de la recherche d'images. Elle comprend quatre ensembles d'images, deux systèmes de recherche d'images et douze prédicteurs de performance de requêtes. Les résultats montrent que problème de la prédiction de performance des requêtes pour la recherche d'images est non résolu car aucun des prédicteurs n'a obtenu une performance élevée pour tous les ensembles de données.

3. https://github.com/naver/deep-image-retrieval

Références

ARABZADEH N., ZARRINKALAM F., JOVANOVIC J., AL-OBEIDAT F. & BAGHERI E. (2020). Neural embedding-based specificity metrics for pre-retrieval query performance prediction. *Information Processing & Management*, **57**(4), 102248.

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

CHEN X., HE B. & SUN L. (2022). Groupwise Query Performance Prediction with BERT. In *Proceedings of ECIR*, p. 64–74.

CHIFU A.-G., LAPORTE L., MOTHE J. & ULLAH M. Z. (2018). Query Performance Prediction Focused on Summarized Letor Features. In *Proceedings of SIGIR*, p. 1177–1180.

CRONEN-TOWNSEND S., ZHOU Y. & CROFT W. B. (2002). Predicting query performance. In *Proceedings of SIGIR*, p. 299–306.

CUMMINS R. (2014). Document score distribution models for query performance inference and prediction. *ACM Transactions on Information Systems*, **32**(1), 2.

CUMMINS R., JOSE J. & O'RIORDAN C. (2011). Improved query performance prediction using standard deviation. In *Proceedings of SIGIR*, p. 1089–1090.

DATTA S., MACAVANEY S., GANGULY D. & GREENE D. (2022). A'pointwise-query, listwise-document'based query performance prediction approach. In *Proceedings of SIGIR*, p. 2148–2153.

DÉJEAN S., IONESCU R. T., MOTHE J. & ULLAH M. Z. (2020). Forward and backward feature selection for query performance prediction. In *Proceedings of SAC*, p. 690–697.

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

EVERINGHAM M., ESLAMI S. A., VAN GOOL L., WILLIAMS C. K., WINN J. & ZISSERMAN A. (2015). The PASCAL Visual Object Classes Challenge : A Retrospective. *International journal of computer vision*, **111**, 98–136.

FAGGIOLI G., ZENDEL O., CULPEPPER J. S., FERRO N. & SCHOLER F. (2022). sMARE : a new paradigm to evaluate and understand query performance prediction methods. *Information Retrieval Journal*, **25**(2), 94–122.

HAUFF C., AZZOPARDI L. & HIEMSTRA D. (2009). The combination and evaluation of query performance prediction methods. In *Proceedings of ECIR*, p. 301–312.

HAUFF C., HIEMSTRA D. & DE JONG F. (2008). A survey of pre-retrieval query performance predictors. In *Proceedings of CIKM*, p. 1419–1420.

HE B. & OUNIS I. (2004). Inferring query performance using pre-retrieval predictors. In *Proceedings of SPIRE*, p. 43–54.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of CVPR*, p. 770–778.

IONESCU R., ALEXE B., LEORDEANU M., POPESCU M., PAPADOPOULOS D. P. & FERRARI V. (2016). How hard can it be ? estimating the difficulty of visual search in an image. In *Proceedings of CVPR*, p. 2157–2166.

JAFARZADEH P. & ENSAN F. (2022). A semantic approach to post-retrieval query performance prediction. *Information Processing & Management*, **59**(1), 102746.

JIA Q. & TIAN X. (2015). Query difficulty estimation via relevance prediction for image retrieval. *Signal Processing*, **110**, 232–243.

JIA Q., TIAN X. & MEI T. (2014). Query difficulty estimation via pseudo relevance feedback for image search. In *Proceedings of ICME*, p. 1–6.

KATZ G., SHTOCK A., KURLAND O., SHAPIRA B. & ROKACH L. (2014). Wikipedia-based query performance prediction. In *Proceedings of SIGIR*, p. 1235–1238.

KURLAND O., RAIBER F. & SHTOK A. (2012). Query-performance prediction and cluster ranking : Two sides of the same coin. In *Proceedings of CIKM*, p. 2459–2462.

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.

LI F.-F., ANDREETO M., RANZATO M. & PERONA P. (2022). Caltech 101. DOI : 10.22002/D1.20086.

LI Y., GENG B., YANG L., XU C. & BIAN W. (2012). Query difficulty estimation for image retrieval. *Neurocomputing*, **95**, 48–53.

MIZZARO S., MOTHE J., ROITERO K. & ULLAH M. Z. (2018). Query performance prediction and effectiveness evaluation without relevance judgments : Two sides of the same coin. In *Proceedings of SIGIR*, p. 1233–1236.

MOTHE J. & TANGUY L. (2005). Linguistic features to predict query difficulty. In *Proceedings of SIGIR*, p. 7–10.

NIE L., WANG M., ZHA Z.-J. & CHUA T.-S. (2012). Oracle in image search : a content-based approach to performance prediction. *ACM Transactions on Information Systems*, **30**(2), 1–23.

PEDRONETTE D. C. G. & TORRES R. D. S. (2015). Unsupervised effectiveness estimation for image retrieval using reciprocal rank information. In *Proceedings of SIBGRAPI*, p. 321–328.

POESINA E., IONESCU R. T. & MOTHE J. (2023). iqpp : A benchmark for image query performance prediction. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

RADENOVIĆ F., ISCEN A., TOLIAS G., AVRITHIS Y. & CHUM O. (2018). Revisiting oxford and paris : Large-scale image retrieval benchmarking. In *Proceedings of CVPR*, p. 5706–5715.

RADENOVIĆ F., TOLIAS G. & CHUM O. (2019). Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**(7), 1655–1668.

RAIBER F. & KURLAND O. (2014). Query-performance prediction : setting the expectations straight. In *Proceedings of SIGIR*, p. 13–22.

REVAUD J., ALMAZÁN J., REZENDE R. S. & SOUZA C. R. D. (2019). Learning with Average Precision : Training Image Retrieval with a Listwise Loss. In *Proceedings of ICCV*, p. 5107–5116.

ROITMAN H. (2018). An extended query performance prediction framework utilizing passage-level information. In *Proceedings of SIGIR*, p. 35–42.

ROITMAN H., ERERA S. & WEINER B. (2017). Robust standard deviation estimation for query performance prediction. In *Proceedings of SIGIR*, p. 245–248.

ROY D., GANGULY D., MITRA M. & JONES G. J. (2019). Estimating Gaussian mixture models in the local neighbourhood of embedded word vectors for query performance prediction. *Information Processing & Management*, **56**(3), 1026–1045.

RUSSAKOVSKY O., DENG J., SU H., KRAUSE J., SATHEESH S., MA S., HUANG Z., KARPATHY A., KHOSLA A., BERNSTEIN M. *et al.* (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, **115**, 211–252.

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. (2010). Using statistical decision theory and relevance models for query-performance prediction. In *Proceedings of SIGIR*, p. 259–266.

SOVIANY P., IONESCU R. T., ROTA P. & SEBE N. (2021). Curriculum self-paced learning for cross-domain object detection. *Computer Vision and Image Understanding*, **204**, 103–166.

SUN S., ZHOU W., TIAN Q., YANG M. & LI H. (2018). Assessing image retrieval quality at the first glance. *IEEE Transactions on Image Processing*, **27**(12), 6124–6134.

TIAN X., JIA Q. & MEI T. (2015). Query difficulty estimation for image search with query reconstruction error. *IEEE Transactions on Multimedia*, **17**(1), 79–91.

TIAN X., LU Y. & YANG L. (2012). Query difficulty prediction for Web image search. *IEEE Transactions on Multimedia*, **14**(4), 951–962.

VALEM L. P. & PEDRONETTE D. C. G. (2021). A denoising convolutional neural network for self-supervised rank effectiveness estimation on image retrieval. In *Proceedings of ICMR*, p. 294–302.

XING X., ZHANG Y. & HAN M. (2010). Query difficulty prediction for contextual image retrieval. In *Proceedings of ECIR*, p. 581–585.

YOM-TOV E., FINE S., CARMEL D. & DARLOW A. (2005). Learning to estimate query difficulty : including applications to missing content detection and distributed information retrieval. In *Proceedings of SIGIR*, p. 512–519.

ZAMANI H., CROFT W. B. & CULPEPPER J. S. (2018). Neural query performance prediction using weak supervision from multiple signals. In *Proceedings of SIGIR*, p. 105–114.

ZHAO Y., SCHOLER F. & TSEGAY Y. (2008). Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Proceedings of ECIR*, p. 52–64 : Springer.