Prédictibilité de la prédiction de la performance des requêtes? Une approche basée sur les valeurs aberrantes multivariées

Adrian-Gabriel Chifu, Sébastien Déjean, Moncef Garouani, Josiane Mothe, Diégo Ortiz, Md Zia Ullah

Prédictibilité de la prédiction de la performance des requêtes ? Une approche basée sur les valeurs aberrantes multivariées

Adrian-Gabriel Chifu[1,*,†], Sébastien Déjean[2,†], Moncef Garouani[3,†], Josiane Mothe[4,†], Diégo Ortiz[5,†] and Md Zia Ullah[6,†]

[1]Aix-Marseille Université, Université de Toulon, CNRS, LIS, France
[2]Université de Toulouse, IMT, URM5219 CNRS, France
[3]Université de Toulouse, UTC, IRIT, UMR5505 CNRS, Toulouse, France
[4]INSPE, Université de Toulouse, UT2J, IRIT, URM5505 CNRS, France
[5]IRIT, UMR5505, Toulouse, France
[6]Edinburgh Napier University, Edinburgh, UK

Abstract

Ceci est un résumé en français de l'article intitulé "Can we predict QPP? An approach based on multivariate outliers", accepté pour la conférence ECIR 2024 dans la catégorie papiers courts [1].

Keywords

Recherche d'Information, Prédiction de la performance des requêtes, QPP, Caractéristiques post-recherche, Détection de valeurs aberrantes, Transformed Rank Correlation

1. Introduction

La prédiction de la performance des requêtes (QPP) représente un défi majeur en recherche d'information (RI). Il s'agit de prédire l'efficacité d'un moteur de recherche pour une variété de requêtes. Malgré les avancées dans ce domaine, la prédiction reste imparfaite [2, 3, 4, 5, 6, 7, 8, 9]. Nous avançons l'hypothèse que la performance de certaines requêtes est intrinsèquement plus difficile à prévoir que pour d'autres. Afin d'identifier ces requêtes, nous utilisons la méthode des Corrélations des Rangs Transformées (TRC pour *Transformed Rank Correlation*), une méthode de détection de valeurs aberrantes multivariées. L'évaluation de cette méthode sur plusieurs collections et avec différents types de prédicteurs de performance montre que la méthode est robuste et améliore la corrélation entre la prédiction de performance et la performance

CORIA 2024 (COnférence en Recherche d'Information et Applications), 3-4 avril 2022, La Rochelle, France
*Auteur correspondant.
†Ces auteurs ont contribué de manière égale.
adrian.chifu@univ-amu.fr (A. Chifu); sebastien.dejean@math.univ-toulouse.fr (S. Déjean); moncef.garouani@irit.fr (M. Garouani); Josiane.Mothe@irit.fr (J. Mothe); diego.ortiz@irit.fr (D. Ortiz); m.ullah@napier.ac.uk (M. Z. Ullah)
https://adrianchifu.com (A. Chifu); https://perso.math.univ-toulouse.fr/dejean/ (S. Déjean); http://www.irit.fr/~Josiane.Mothe (J. Mothe)
0000-0003-4680-5528 (A. Chifu); 0000-0001-9610-5306 (S. Déjean); 0000-0003-2528-441X (M. Garouani); 0000-0001-9273-2193 (J. Mothe); 0009-0002-7272-5644 (D. Ortiz); 0000-0002-4022-7344 (M. Z. Ullah)

constatée. La méthode TRC prend en compte l'interaction entre plusieurs variables prédicteurs de performance. Nous l'utilisons pour identifier les requêtes présentant des prédictions anormalement dispersées. Elle calcule la distance, basée sur des corrélations des rangs, entre chaque observation et la tendance centrale des données. Les observations dont la distance dépasse un seuil (valeur par défaut fixée au quantile 0.95) sont identifiées comme des valeurs aberrantes.

2. Données et Résultats

Nous utilisons les collections ad-hoc de TREC[1], spécifiquement TREC78 et WT10G, et deux mesures d'évaluation de performance des moteurs de recherche d'information: la précision moyenne (AP@1000) et le gain cumulatif normalisé décroissant (nDCG@20). Nous analysons successivement deux groupes de predicteurs : d'une part des caractéristiques LETOR: LemurTF_IDF, In_expC2, InB2, et InL2 initialement conçues pour le ré-ordonnancement de documents [10] et agrégés comme proposé dans [11] et d'autre part quatre prédicteurs de la littérature: Normalized query commitment (NQC) [12], Unnormalized query commitment (UCQ) [12], QF (query feedback) [13] et WIG [13]. Les quatre caractéristiques que nous avons conservées sont `LemurTF_IDF`, `In_expC2`, `InB2` et `InL2`, agrégées en utilisant la fonction maximum. Notre analyse a mis en évidence une amélioration notable de la corrélation entre les prédictions et les performances réelles des moteurs de recherche sur les ensembles de requêtes étudiés. Ceci est particulièrement vrai sur la collection TREC78 où, après l'exclusion des requêtes identifiées par notre méthode, la corrélation de Pearson pour le prédicteur `LemurTF_IDF` augmente de façon significative de 0.522 à 0.700. Cela souligne non seulement l'efficacité de notre approche, mais aussi son potentiel pour affiner la précision des prédictions de performance des requêtes. Les requêtes identifiées comme difficiles à prédire par notre méthode présentaient souvent des caractéristiques complexes, telles que l'ambiguïté de la requête ou une concordance limitée avec les documents pertinents de la collection. Cette observation indique que notre méthode ne se limite pas à une simple identification statistique des requêtes aberrantes, mais qu'elle est également capable de refléter des défis plus profonds liés à la prédiction de performance dans des contextes RI variés. D'après ces premiers résultats, notre méthode semble robuste. Elle permet une compréhension plus nuancée des interactions entre les requêtes, les documents et les prédicteurs de performance, et devrait offrir un moyen pour améliorer la conception et l'ajustement des modèles de prédiction de performance dans le cadre de la RI.

3. Conclusion

Cette étude montre la faisabilité et l'efficacité de l'identification des requêtes difficiles à prédire à l'aide de la détection de valeurs aberantes s'appuyant sur plusieurs variables, en particulier via l'utilisation de la méthode TRC. En écartant les requêtes identifiées comme problématiques, nous améliorons la précision globale des prédictions de performance des requêtes, ouvrant ainsi de nouvelles perspectives pour la recherche dans ce domaine. Nos recherches futures viseront à explorer l'application de cette méthode à d'autres ensembles de données ainsi qu'à des modèles combinant plusieurs prédicteurs.

[1] trec.nist.gov: Text REtrieval Conference

References

[1] A.-G. Chifu, S. Déjean, G. Garouani, Moncef, J. Mothe, D. Ortiz, M. Z. Ullah, Can we predict QPP? An approach based on multivariate outliers, in: European Conference on Information Retrieval, Springer, 2024.

[2] S. Cronen-Townsend, Y. Zhou, W. B. Croft, Predicting query performance, in: International ACM SIGIR conference on Research and development in information retrieval, 2002, pp. 299–306.

[3] G. Amati, C. Carpineto, G. Romano, Query difficulty, robustness, and selective application of query expansion, in: S. McDonald, J. Tait (Eds.), Advances in Information Retrieval, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 127–137.

[4] C. Hauff, D. Hiemstra, F. de Jong, A survey of pre-retrieval query performance predictors, in: ACM CIKM, 2008, pp. 1419–1420.

[5] D. Carmel, E. Yom-Tov, Estimating the query difficulty for information retrieval, Morgan & Claypool Publishers, 2010.

[6] F. Raiber, O. Kurland, Query-performance prediction: Setting the expectations straight, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14, ACM, New York, NY, USA, 2014, pp. 13–22. URL: http://doi.acm.org/10.1145/2600428.2609581. doi:10.1145/2600428.2609581.

[7] R. Deveaud, J. Mothe, M. Z. Ullah, J.-Y. Nie, Learning to adaptively rank document retrieval system configurations, ACM Transactions on Information Systems (TOIS) 37 (2018) 3. doi:10.1145/3231937.

[8] S. Datta, S. MacAvaney, D. Ganguly, D. Greene, A'pointwise-query, listwise-document'based query performance prediction approach, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2148–2153.

[9] G. Faggioli, T. Formal, S. Marchesin, S. Clinchant, N. Ferro, B. Piwowarski, Query performance prediction for neural ir: Are we there yet?, in: European Conference on Information Retrieval, Springer, 2023, pp. 232–248.

[10] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, H. Li, Learning to rank: from pairwise approach to listwise approach, in: ICML, 2007, pp. 129–136.

[11] A.-G. Chifu, L. Laporte, J. Mothe, M. Z. Ullah, Query performance prediction focused on summarized letor features, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 1177–1180.

[12] A. Shtok, O. Kurland, D. Carmel, F. Raiber, G. Markovits, Predicting query performance by query-drift estimation, ACM Transactions on Information Systems (TOIS) 30 (2012) 11.

[13] Y. Zhou, W. B. Croft, Query performance prediction in web search environments, in: ACM SIGIR, 2007, pp. 543–550.