Outil d'exploration visuelle de l'évaluation longitudinale en Recherche d'Information

Gabriela Gonzalez-Saez[1], Petra Galuščáková[2], Romain Deveaud[3], Philippe Mulhem[1] and Lorraine Goeuriot[1]

[1]*Univ. Grenoble Alpes, CNRS, Grenoble INP[i], LIG, 38000 Grenoble, France*
[2]*University of Stavanger, Stavanger, Norway*
[3]*Qwant, Paris, France*

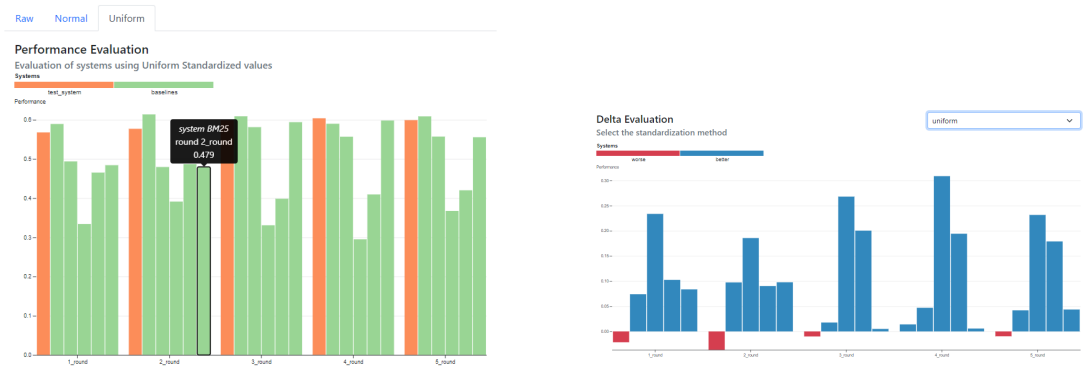
1. Description

Nous présentons un outil *open source* destinée aux chercheurs et ingénieurs travaillant sur les moteurs de recherche, qui évaluent de manière répétée les systèmes de recherche d'information (SRI). Nous visons à faciliter l'évaluation continue de ces systèmes et à identifier leurs améliorations possibles. Notre outil répond à ces besoins en proposant un ensemble de visualisations qui présentent les changements de performances des systèmes au cours des évolutions des collections de test. Les résultats des SRI dépendent des corpus de documents des requêtes et des pertinences utilisateurs. Nous proposons un tableau de bord interactif et dynamique qui permet aux utilisateurs de suivre de manière continue leur système au cours de l'évolution de collections de tests (*rounds*). L'état de l'art, comme Vis-Trec [1], RecDelta [2], ou DiffIR [3], ne prend pas en compte l'évolution de l'évaluation des systèmes au cours de rounds. Une évaluation continue est basée sur l'évaluation répétée, sur des rounds, des systèmes. Notre interface présente des visualisations qui affichent comment l'évaluation de SRI évolue au cours des rounds, en se basant sur deux méthodes d'évaluation ; la standardisation et le méta-analyse. Les scores de standardisation permettent de tenir compte de la difficulté d'une requête [4]. Webber et al. [5] a proposé des transformations non-linéaires qui supposent une distribution normale des évaluations brutes par requête, alors que Sakai [4] a étudié l'utilisation de transformations linéaires des scores initiaux en supposant une distribution uniforme des évaluations brutes. Notre interface présente les résultats brutes et standardisés, au travers l'utilisation d'une fonctions cumulée (*cdf*) de normalisations. Soboroff [6] a proposé d'utiliser une méta-analyse pour évaluer un système sur plusieurs collections en comparant une référence au système évalué par un delta, afin de déterminer une différence moyenne associée à un intervalle de confiance. Nous intégrons ces éléments dans notre interface. Le code, développé avec Django et la librairie D3.js, ainsi que la démonstration, sont accessibles sur: https://github.com/gabrielanicole/ExCEIR.

[i]Institute of Engineering Univ. Grenoble Alpes

CORIA 2024, La Rochelle

gabriela-nicole.gonzalez-saez@univ-grenoble-alpes.fr (G. Gonzalez-Saez); galuscakova@gmail.com (P. Galuščáková); r.deveaud@qwant.com (R. Deveaud); philippe.mulhem@imag.fr (P. Mulhem); lorraine.goeuriot@univ-grenoble-alpes.fr (L. Goeuriot)



(a) Vue globale.

(b) Vue par deltas.

Figure 1: Vue par *rounds* des performances d'un système comparé à 5 références, sur les 5 rounds de TREC-COVID.

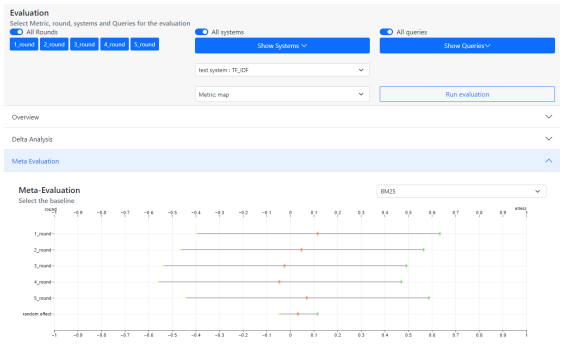


Figure 2: Méta-analyse du système TF_IDF par rapport à la référence BM25.

L'interface peut présenter la vue globale d'un système au cours des rounds. Chaque barre représente la valeur moyenne sur des requêtes et des rounds sélectionnés. Nous avons chois ici toutes les requêtes et tous les rounds sur la figure 1. La performance du système testé est en orange et les systèmes de références sont en vert. Comme décrit plus haut, nous pouvons utiliser les scores d'évaluation bruts ou bien standardisés sur les références. La figure 1a présente les résultats standardisés suivant une loi uniforme. La figure 1b présente les mêmes résultats sous forme de deltas. La figure 2 présente un affichage par forêt, utilisé par [6] pour présenter des résultats de Méta-Analyse : chaque ligne représente l'effet du système testé (ici TF-IDF) par rapport à la référence (ici BM25). Le première ligne montre donc que le système testé est en moyenne meilleur de 0,12 en MAP que le système de référence. L'intervalle dont le maximum est en vert correspond l'intervalle de confiance à 95%.

Tous ces éléments visent donc à permettre d'estimer les capacité des systèmes de recherche d'information à se confronter aux collections évolutives.

Ce travail a été partiellement financé par le projet Kodicare, ANR-19-CE23-0029, de l'Agence Nationale de la Recherche.

References

[1] M. Tamannaee, N. Arabzadeh, E. Bagheri, Vis-trec: A system for the in-depth analysis of trec_eval results, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 2181–2184. URL: https://doi-org.ins2i.bib.cnrs.fr/10.1145/3397271.3401412. doi:10.1145/3397271.3401412.

[2] Y.-S. Chiang, Y.-Z. Liu, C.-F. Tsai, J.-K. Lou, M.-F. Tsai, C.-J. Wang, Recdelta: An interactive dashboard on top-k recommendation for cross-model evaluation, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 3224–3228.

[3] K. M. Jose, T. Nguyen, S. MacAvaney, J. Dalton, A. Yates, Diffir: Exploring differences in ranking models' behavior, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2595–2599.

[4] T. Sakai, A simple and effective approach to score standardisation, in: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval, 2016, pp. 95–104.

[5] W. Webber, A. Moffat, J. Zobel, Score standardization for inter-collection comparison of retrieval systems, in: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp. 51–58.

[6] I. Soboroff, Meta-analysis for retrieval experiments involving multiple test collections, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 713–722.