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

Vers l'évaluation continue des systèmes de recherche d'information.

Petra Galuščáková[1] Romain Deveaud[1] Gabriela Gonzalez-Saez[1] Philippe Mulhem[1] Lorraine Goeuriot[1] Florina Piroi[3] Martin Popel[4]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP[*], LIG Grenoble, France
(2) Qwant, France
(3) RSA, Autriche
(4) Charles University, Prague, République Tchèque
Philippe.Mulhem@imag.fr

RÉSUMÉ

Cet article présente le corpus de données de la campagne d'évaluation LongEval, dans le cadre de CLEF 2023. L'objectif de cette campagne est d'étudier comment les systèmes de recherche d'informations réagissent aux changements des données qu'ils traitent, en particulier les documents et les requêtes. Nous détaillons les objectifs de la tâche, le processus d'acquisition de données et les mesures d'évaluation utilisés.

ABSTRACT

Toward continuous evaluation of Web Information Retrieval.

This article provides a brief overview of the LongEval evaluation campaign's data corpus, which is part of CLEF 2023. The aim of this campaign is to investigate how information retrieval systems respond to changes in the data they process, specifically documents and queries. We detail the task objectives, data acquisition process, and evaluation metrics used in the campaign.

MOTS-CLÉS : Collection de test, Recherche sur le Web.

KEYWORDS: Test collection, Web search.

1 Introduction

(Ren *et al.*, 2022 arxiv220412755) a démontré que la qualité d'un système de Recherche d'Information (SRI) neuronal profond dépend de la cohérence entre les données d'entraînement et de test. Cependant, il n'existe pas à notre connaissance de campagne d'évaluation dédié à cette étude. La tâche de recherche d'information (RI) de la campagne LongEval 2023 vise à évaluer le comportement des systèmes de recherche d'information modernes face à l'évolution des données. Plus précisément, cette tâche vise à mieux comprendre l'impact du temps sur les systèmes de RI afin : i) d'évaluer l'efficacité des différentes approches de recherche d'information dans le temps, et ii) de proposer des modèles et systèmes de RI capables de tirer parti des ensembles d'apprentissages vieillissants, tout en minimisant la baisse des performances au fil du temps. La figure 1 présente le processus d'évaluation proposé dans le cadre de cette tâche. Au temps *t*, un système utilise des données d'apprentissage

∗. Institute of Engineering Univ. Grenoble Alpes.

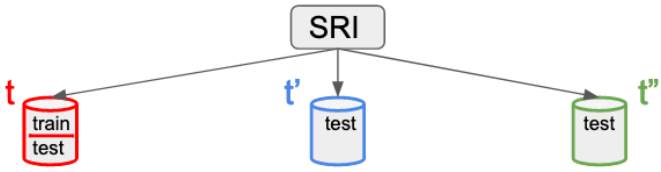

FIGURE 1 – La tâche RI de LongEval apprentissage à t, tests aux temps t, t' et t".

(documents, requêtes, évaluation de pertinence), noté *train*. A *t*, un ensemble additionnel de requêtes, de *test*, est fourni sur ces mêmes documents. L'évaluation du système sur ces requêtes de test par les organisateurs de la campagne donne la mesure de référence du système. Pour les temps ultérieurs *t'* et *t"*, qui suivent *t*, un nouveau corpus de documents et un nouvel ensemble de requêtes de test sont fournis, suivis d'une évaluation par les organisateurs de la campagne. On mesure ensuite la dégradation des résultats entre la référence de test et les temps *t'* et *t"*. En section 2, nous présentons les objectifs de la tâche de recherche d'information de la campagne LongEval. La section 3 présente les données fournies et leur acquisition. La section 4 détaille les mesures d'évaluation proposées et la section 5 des expérimentations préliminaires. Nous concluons en partie 6.

2 Principes et objectifs

La collection de la tâche de recherche d'information de Longeval s'appuie sur un large ensemble de données (un corpus de 4,2 missions de pages, 2500 de requêtes réelles, un grand nombre d'évaluations de pertinence tirées de vraies interactions avec des utilisateurs) fourni par le moteur de recherche Qwant (`https://www.qwant.com`). Elle reflète les changements de la recherche de pages Web à travers le temps, en fournissant un corpus de documents et des requêtes qui évoluent. A notre connaissance, de telles caractéristiques ne sont pas proposées dans une campagne d'évaluation, à l'échelle sur laquelle nous nous concentrons. Le paradigme de Cranfield, classique dans l'évaluation de la recherche d'information, ne prend pas en compte d'élément temporel : le corpus, les requêtes, et les évaluations de pertinence sont fixés une fois pour toute. Les collections de test Robust (Voorhees, 2006) et Twitter (Sequiera & Lin, 2017) contiennent les dates de création des documents, mais ne considèrent pas cette information en tant qu'objet d'étude. La seule collection qui intègre explicitement l'aspect temporel dans la recherche d'information ad-hoc est le récent jeu de données TREC-COVID (Voorhees *et al.*, 2021), qui propose un corpus évolutif sur le COVID. TREC-COVID contient quelques dizaines de milliers de documents et 45 requêtes au total. Dans LongEval, les échelles de données sur bien supérieures : LongEval est la seule grande collection avec des données acquises en 2022, qui a pour objectif d'évaluer les capacités des systèmes de RI modernes à se confronter à des données évolutives. De plus, le corpus de LongEval est en français et en anglais, ce qui nous différencie des collections multilingues telles que CLEF eHealth (Kelly *et al.*, 2016, 2019), dans lesquelles le français et l'anglais sont intégrées de manière limitée.

3 Les données

Nous décrivons ici le processus général d'acquisition des données issues du moteur de recherche Web Qwant, et la création des différents composants de la collection. L'acquisition globale est périodique et récurrente dans le temps afin de constituer une séquence de *sous-collections* aux temps t, t' et

t". Une sous-collection présente les caractéristiques d'une collection de test traditionnelle (requêtes, documents, jugements de pertinence), sauf qu'elle partage un ensemble commun de *sujets* (qui ne sont pas les requêtes elles-mêmes) avec les autres sous-collections. Elle est définie par :

1. L'acquisition d'un ensemble de **sujets**, sélectionnés à partir du Web et des médias sociaux. Cette acquisition est basée sur des sujets populaires - mais stables à long terme - et est effectuée une seule fois pour l'ensemble de la collection de recherche d'information LongEval.
2. La sélection de **requêtes de recherche** liées aux sujets ci-dessus, provenant des requêtes réelles émises par les utilisateurs du moteur de recherche Qwant. Elle est basée sur des intersections de chaînes de caractères entre sujets et requêtes.
3. Les **estimations de pertinence**. Nous nous appuyons ici sur deux manières de collecter ces évaluations : en utilisant implicitement des modèles de clics (Chuklin *et al.*, 2015) calculés à partir des logs de requêtes Qwant. Nous intègrerons également des évaluations manuelles qui seront collectées à la suite des runs des participants à la tâche. Étant donné que chaque sous-collection peut contenir plusieurs milliers de requêtes, nous effectuerons des évaluations explicites sur un sous-ensemble de requêtes sélectionnées manuellement.
4. L'acquisition du **corpus de documents**. Ce corpus, fourni par Qwant, est une union de : i) tous les documents Web qui ont été affichés dans la première page de résultats pour chaque requête d'une sous-collection, et ii) un échantillon aléatoire assez important de l'index Qwant. Ce protocole conduit à un corpus qui contient un mélange de documents pertinents et non pertinents. Le processus présenté gère l'évolution des pages Web, car le corpus n'est pas seulement composé d'URL, mais également du contenu des pages Web acquises à un instant.

Ces données sont acquises initialement français. Afin de permettre une participation plus ouverte à des équipes non-francophones, nous proposons une version anglaise de toutes des données (documents et requpêtes). Pour la traduction, nous avons utilisé le système français-anglais CUBBITT (Popel *et al.*, 2020), disponible à `https://lindat.cz/services/translation`. Suivant ce principe, on acquiert au temps *t*, *t'* et *t"* des corpus complets, en utilisant le même ensemble de sujets. Les requêtes de *train* au temps *t* sont sont fournies aux participants avec leurs évaluations de pertinence, et les requêtes de test sans évaluations de pertinence. Ces éléments, cf. section 4, sont utilisés par l'évaluation. Au temps *t'* et *t"*, seuls des ensembles de tests sont fournis. Les éléments du processus permettent de répondre aux objectifs de la tâche pour les raisons suivantes : a) Les corpus de documents que nous utilisons proviennent de la même source, Qwant, suivant le même principe. Ces corpus reflètent donc bien une évolution temporelle ; b) les requêtes sont tirées d'un ensemble de sujets stables dans le temps, ce qui permet d'éviter a priori des biais sur des sujets incomparables en terme de comportement humain ; c) Les évaluations de pertinence, calculées automatiquement à partir d'interactions d'utilisateurs, permettent de pouvoir gérer de grande quantités de requêtes.

4 Les mesures d'évaluation

Nous utilisons : Le nDCG, qui permet de bien prendre en compte l'importance des positions dans la liste de réponses ainsi que des mesures de pertinence multivaluées. ERR (Chapelle *et al.*, 2009) : nos estimations de pertinence étant calculées à base de Click Models, nous pouvons également considérer la métrique Expected Reciprocal Rank qui suit un modèle de navigation de l'utilisateur similaire à celui de la nDCG tout en utilisant des probabilités d'attractivité des documents. RnD(t,t') et RnD(t,t'') : la chute relative du nDCG. Ces valeurs sont égales à la différence entre le nDCG sur

les données de test au temps t et sur les ensembles de tests à t' et t''. Elles quantifient la robustesse du système évalué par rapport aux évolutions des corpus de documents et des requêtes. Avec ces mesures, un SRI de bonne qualité et répondant bien à l'évolution obtiendra une valeur de nDCG et de ERR élevée à t, t' et t'', ainsi qu'une valeur élevée pour le RnD(t,t') et RnD(t,t'').

5 Résultats sur l'ensemble d'apprentissage au temps *t*

Nous présentons dans le tableau 1 des résultats préliminaires (mesures classiques et ERR@20) sur les données d'apprentissage en français au temps t, sans apprentissage spécifique. Nous avons testé le système Terrier (Macdonald *et al.*, 2012) fournissant 1000 documents par requête, avec les paramètres par défaut (anti-dictionnaire français [1] ; troncature sur le français, *FrenchSnowballStemmer*), ainsi qu'un *reranking* de ce même BM25 par le modèle monoT5 (Nogueira *et al.*, 2020) de Castorini accessible par Pygaggle, réglé finement sur MS MARCO v1. Ces résultats nous permettent de vérifier que le corpus proposé fonctionne correctement avec des modèles classiques, tout en présentant une forte marge de progression, comme on le voit avec un *reranking* simple basé sur T5.

Système	P@10	nDCG@10	ERR@20	MAP	nDCG	Reciprocal Rank
BM25	0,1109	0,2083	0,0379	0,1767	0,3308	0,3019
reranking T5	0,1329	0,2578	0,0460	0,2175	0,3308	0,3578

TABLE 1 – Évaluation du modèle BM25 de Terrier, et reranking top 100 par T5, sur les requêtes d'apprentissage au temps *t*.

6 Conclusion

Dans cet article, nous avons décrit le corpus d'évaluation LongEval, destiné à évaluer dans quelle mesure les systèmes modernes de RI se comportent face à l'évolution des données dans le cadre de la recherche de documents sur le Web. Nous avons détaillé les étapes d'acquisition des données en françsi, leur traduction en anglais, les mesures d'évaluation et quelques résultats utilisant un modèle classiques BM25. Le corpus est accessible via le site de Longeval https://clef-longeval.github.io/.

Remerciements

Ce travail est soutenu par le projet bilatéral ANR Kodicare, subvention ANR-19-CE23-0029 de l'Agence Nationale de la Recherche française, et par le Fonds scientifique autrichien (FWF, subvention I4471-N). Ce travail a utilisé l'infrastructure de recherche LINDAT/CLARIAH-CZ (https ://lindat.cz), soutenue par le ministère de l'Éducation, de la Jeunesse et des Sports de la République tchèque (projet LM2018101 et projet LM2023062).

1. https://www.kaggle.com/datasets/rtatman/stopword-lists-for-19-languages?select=frenchST.txt

Références

CHAPELLE O., METLZER D., ZHANG Y. & GRINSPAN P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, p. 621–630, New York, NY, USA : Association for Computing Machinery.

CHUKLIN A., MARKOV I. & RIJKE M. D. (2015). Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, **7**(3), 1–115.

KELLY L., GOEURIOT L., SUOMINEN H., NÉVÉOL A., PALOTTI J. & ZUCCON G. (2016). Overview of the CLEF eHealth evaluation lab 2016. In N. FUHR, P. QUARESMA, T. GONÇALVES, B. LARSEN, K. BALOG, C. MACDONALD, L. CAPPELLATO & N. FERRO, Éds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, p. 255–266, Cham : Springer International Publishing.

KELLY L., SUOMINEN H., GOEURIOT L., NEVES M., KANOULAS E., LI D., AZZOPARDI L., SPIJKER R., ZUCCON G., SCELLS H. & PALOTTI J. (2019). Overview of the CLEF eHealth evaluation lab 2019. In F. CRESTANI, M. BRASCHLER, J. SAVOY, A. RAUBER, H. MÜLLER, D. E. LOSADA, G. HEINATZ BÜRKI, L. CAPPELLATO & N. FERRO, Éds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, p. 322–339, Cham : Springer International Publishing.

MACDONALD C., MCCREADIE R., SANTOS R. L. & OUNIS I. (2012). From puppy to maturity : Experiences in developing terrier. *Proc. of OSIR at SIGIR*, p. 60–63.

NOGUEIRA R., JIANG Z., PRADEEP R. & LIN J. (2020). Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 708–718, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.63.

POPEL M., TOMKOVA M., TOMEK J., ŁUKASZ KAISER, USZKOREIT J., BOJAR O. & ŽABOKRTSKÝ Z. (2020). Transforming machine translation : a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, **11**(4381), 1–15.

REN R., QU Y., LIU J., ZHAO W. X., WU Q., DING Y., WU H., WANG H. & WEN J.-R. (2022, arxiv :2204.12755). A thorough examination on zero-shot dense retrieval. DOI : 10.48550/ARXIV.2204.12755.

SEQUIERA R. & LIN J. (2017). Finally, a downloadable test collection of tweets. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, p. 1225–1228, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3077136.3080667.

VOORHEES E., ALAM T., BEDRICK S., DEMNER-FUSHMAN D., HERSH W. R., LO K., ROBERTS K., SOBOROFF I. & WANG L. L. (2021). Trec-covid : Constructing a pandemic information retrieval test collection. *SIGIR Forum*, **54**(1). DOI : 10.1145/3451964.3451965.

VOORHEES E. M. (2006). The trec 2005 robust track. In *ACM SIGIR Forum*, volume 40, p. 41–48 : ACM New York, NY, USA.