Team OpenWebSearch at CLEF 2024: QuantumCLEF

Maik **Fröbe**[1], Daria **Alexander**[2], Gijs **Hendriksen**[2], Ferdinand **Schlatt**[1], Matthias **Hagen**[1] and Martin **Potthast**[3]

[1]*Friedrich-Schiller-Universität Jena*

[2]*Radboud Universiteit Nijmegen*

[3]*University of Kassel, hessian.AI, ScaDS.AI*

Abstract

We describe the OpenWebSearch group's participation in the CLEF 2024 QuantumClef IR Feature Selection track. Our submitted runs focus on the observation that the importance of features in learning-to-rank models can vary and contradict itself when changing the training setup. To address this problem and identify a subset of features that is robust across diverse downstream training procedures, we bootstrap feature importance scores by repeatedly training models on randomly selected subsets of features and measuring their importance in trained models. We indeed observe that feature importance varies widely across different bootstraps and also contradicts itself. We hypothesized that quantum annealers could better explore this complex optimization landscape than simulated annealers. However, we find that quantum annealers do not find substantially more optimal solutions that yield substantially more effective learning-to-rank models.

Keywords

learning-to-rank, bootstrapping, feature selection

1. Introduction

Learning-to-Rank aims to identify a combination of features that produce an effective ranking [1]. Even in the era of pre-trained transformers [2], feature-based learning-to-rank remains important as it can integrate features not available in transformers, compensating for knowledge to which transformers have no access [3, 4]. Especially commercial search engines might combine many features, e.g., a recent leak claims that Google search incorporates more than 14 000 features into their ranking.[1]

Such scenarios highlight the importance of proper feature selection, as different search systems (even if they might be bundled behind a single UI) might target at different tasks (expressed via an evaluation scenario, e.g., evaluation measure with a test dataset) that require different sets of features. In the scenario of the QuantumCLEF task [5, 6, 7], we start from the original quadratic unconstrained binary optimization prepared in the official tutorial [8] and contrast the components of this optimization problem with bootstrapped alternatives. Bootstrapping is a frequently used approach in statistics if the mean of some population is not meaningful or can not be calculated (e.g., for categorical values) that draws repeated samples of some data [9]. We use bootstrapping for feature selection by repeatedly sampling LambdaMART models from the training data. Thereby, we follow the intuition that the original optimization problem that uses the mutual information and the conditional mutual information can not capture all potentially interesting dependencies that might impact what features are important. Our code and the bootstrapped feature-importance scores are available online.[2]

2. Related Work

We will review related work on bootstrapping and feature selection in information retrieval that inspired our work.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[1]https://sparktoro.com/blog/an-anonymous-source-shared-thousands-of-leaked-google-search-api-documents-with-me-everyone-in-seo-should-see-them

[2]https://bitbucket.org/eval-labs/qc24-ows/

Algorithm 1 Bootstrapping Feature Importance Scores

Require:

         f, y features for learning to rank with target predictions y

           b number of desired bootstrapped feature importance scores

      lightGBM   lightGBM training procedure

       sample   a sampling approach

1:   $X \leftarrow []$
2:   **while** $i \leq b$ **do**
3:      $f', y' \leftarrow$ sample(f, y)
4:      model $\leftarrow$ lightGBM.train(f', y')
5:      $X \leftarrow X + [$model.calculateFeatureImportance$()]$
6:   **end while**
7:   **return** X

Bootstrapping in Information Retrieval   Bootstrapping, i.e., the process of repeatedly sampling from the same distribution, has been used previously in information retrieval, e.g., to sample from the relevance judgments, from the topics, or from the document corpus [10]. The leave-out-uniques test is a form of re-sampling of relevance judgments used to estimate the reusability of test collections [11, 12, 13]. Bootstrapping topics has been used for significance tests [14, 15] respectively for assessing the discriminatory power of evaluation measures [16, 17, 18]. Analogously, bootstrapping the document corpus can help to simulate different corpora [18], estimate if results transfer to other corpora [19], or, again, for meta evaluations of evaluation measures [18]. Given the wide applicability of bootstrapping in the field of information retrieval, we now intend to apply it to learning to rank. Contrary to the approaches discussed above, our approach mainly focuses on re-sampling the set of features that subsequent learning-to-rank models can access.

Feature Selection   Feature selection approaches are either filter methods, wrapper methods, or embedded methods [20], distinguished on how deep (if at all) they integrate with the learning algorithm [21]. Filter methods have no integration with the learning algorithm [21] (i.e., they run before the learning starts), e.g., the original quadratic unconstrained binary optimization prepared in the official QuantumCLEF tutorial [8] falls into this category. Wrapper methods use a search algorithm to select the features [22], whereas embedd methods integrate the selection into the actual learning phase [21]. Our approach falls into the category of wrapper methods. There is already an high number of existing feature selection approaches for learning to rank [22, 21, 23, 24, 25, 26], comparing respectively integrating them with boostrapping could be interesting directions for future work.

3. Selecting Important Features with Bootstrapping

This section describes our bootstrapping approach for feature selection. Conceptually, we formulate a quadratic unconstrained binary optimization problem [5] that can be optimized via simulated annealing and via quantum annealing. The number of features that our feature selection selects is an hyperparameter that one could optimize, but we leave this for future work and always select the top-25 features (our focus was on the MQ2007 dataset that had around 50 features, so we intuitively selected 25 as number of features to target at). We create three optimization formulations for our bootstrapping feature selection that differ in if they incorporate mutual information optimization objectives or not. We submitted our three approaches within the qCLEF platform [27] for simulated annealers and quantum annealers, yielding 6 runs overall.

     Algorithm 1 shows our bootstrapping algorithm. The algorithm has the features f, the target label y, the number of bootstraps b, an LightGBM training procedure, and a sampling approach as input. Subsequently, each bootstrapping iteration first samples a subset of features f' together with their corresponding ground truth labels y'. With this sampled set of features, a LambdaMART model is

trained for which the feature importance is calculated and added to the return vector X. For the training of the LambdaMART models, we use the LightGBM [28] implementation in PyTerrier [29]. We do not tune the hyperparameters of LambdaMART but use the hyperparameters from a different project without adoption [30]. We sample the featured f' by randomly sorting the feature records and selecting a random subset of 25 features.

To incorporate the bootstrapped feature importance scores into the feature selection, we include them into an optimization criterion that can be optimized by quantum annealers and by simulated annealing. Therefore, we use the quadratic unconstrained binary optimization (QUBO) formulation that minimizes the following objective [5]:

$$\vec{x}^T \cdot Q \cdot \vec{x} = \sum_i^N q_i \cdot x_i + \sum_{i<j}^N q_{i,j} \cdot x_i \cdot x_j$$

Where $\sum_i^N q_i \cdot x_i$ is the linear part of the QUBO and $\sum_{i<j}^N q_{i,j} \cdot x_i \cdot x_j$ is the quadratic part. The official starting point of the shared task fills the linear part of the QUBO with the negative mutual information between a feature and the ground truth label and the quadratic part with the negative conditional mutual information between two features and the ground truth label [8]. To incorporate our bootstrapped feature importance, we use the following formulation for the linear part:

$$q_i \cdot x_i = \sum_{k=1}^b \frac{X_i^k}{|X|}$$

Where b is the number of bootstraps, X_i^k is the importance of feature i in the k-th bootstrapped model, and $|X|$ is the overall importance. Analogously, we implement the quadratic part of the bootstrapping QUBO via:

$$q_{i,j} \cdot x_i \cdot x_j = \sum_{k=1}^b \frac{X_i^k + X_j^k}{|X|}$$

Where b is the number of bootstraps, X_i^k is the importance of feature i in the k-th bootstrapped model, X_j^k is the importance of feature j in the k-th bootstrapped model, and $|X|$ is the overall importance. In both bootstrapping equations, we skip for a feature i or a feature combination i, j bootstraps that do not include the feature because it was not sampled.

To summarize the points above, we have four parts to build QUBO formulations, two from the original mutual information formulation, and two from our new bootstrapping formulation. We combine them to produce three systems that we run on simulated and quantum annealing:

mi-linear-bootstrapped-boost-3 This QUBO uses the linear part of our bootstrapping formulation and the quadratic part from the original conditional mutual information. We multiple the bootstrapping scores with 3 as this factor provided results on a similar scale then the previous mutual information (identified by manual inspection).

mi-linear-and-quadratic-bootstrapped-boost-3 This QUBO uses the linear and quadratic part of our bootstrapping formulation. We multiple the bootstrapping scores with 3 as this factor provided results on a similar scale then the previous mutual information (identified by manual inspection).

mi-bootstrap-mixture This QUBO uses the average of the mutual information and our bootstrapping variant for the linear and quadratic part.

4. Results

We provide evaluations of our methods compared to the baseline of using all features on the MQ2007 and Istella [3] dataset. We report the results in terms of nDCG@10, reporting the 25-th, the 50-th,

Table 1
Effectiveness of the LambdaMART models in terms of nDCG@10 on the MQ2007 dataset. We report the results of our three feature selection approaches that selected 25 features and their effectiveness at the 25-th, the 50-th and the 75-th percentile and the mean for simulated annealing and quantum annealing.

Feature Selection	Simulated Annealing				Quantum Annealing			
	$\eta_{.25}$	$\eta_{.50}$	$\eta_{.75}$	Mean	$\eta_{.25}$	$\eta_{.50}$	$\eta_{.75}$	Mean
mi-bootstrap-mixture	0.114	0.469	0.727	0.448	0.130	0.474	0.722	0.450
mi-linear-and-quadratic-bootstrapped-boost-3	0.126	0.460	0.733	0.452	0.130	0.450	0.726	0.451
mi-linear-bootstrapped-boost-3	0.118	0.464	0.718	0.451	0.145	0.444	0.716	0.448
Baseline All Features	—	—	—	0.447	—	—	—	—

Table 2
Effectiveness of the LambdaMART models in terms of nDCG@10 on the Istella dataset. We report the results of our three feature selection approaches that selected 25 features and their effectiveness at the 25-th, the 50-th and the 75-th percentile and the mean for simulated annealing and quantum annealing.

Feature Selection	Simulated Annealing				Quantum Annealing			
	$\eta_{.25}$	$\eta_{.50}$	$\eta_{.75}$	Mean	$\eta_{.25}$	$\eta_{.50}$	$\eta_{.75}$	Mean
mi-bootstrap-mixture	0.533	0.681	0.813	0.657	0.491	0.645	0.784	0.621
mi-linear-and-quadratic-bootstrapped-boost-3	0.529	0.677	0.809	0.654	0.473	0.634	0.768	0.609
mi-linear-bootstrapped-boost-3	0.474	0.630	0.772	0.609	0.504	0.655	0.793	0.632
Baseline All Features	—	—	—	0.715	—	—	—	—

and the 75-th quantile ($\eta_{.25}$, $\eta_{.50}$, respectively $\eta_{.75}$) and the Mean of the nDCG@10 for all our three approaches for simulated annealing and quantum annealing.

Table 1 shows the results for the MQ2007 dataset. We observe that all feature selection approaches slightly improve upon the baseline of selecting all features, with the bootstrapping variants outperforming the mixed variant and the QUBO that uses the linear and quadratic bootstrapping part is the most effective one, for simulated and quantum annealing.

Table 2 shows the results for the Istella dataset. We observe that all feature selection approaches are substantially less effective then the baseline of using all features. It is interesting future work to investigate how this can be resolved.

5. Conclusion

We presented the Open Web Search (OWS) team's submission to the QuantumCLEF shared task at CLEF 2023. The motivation behind our approach was that LambdaMART models trained on shuffled datasets might choose different features as important ones. Therefore, we repeatedly train LambdaMART models on randomized feature sets and measure the importance of the features in the trained model. For the MQ2007 dataset, our approach substantially outperforms the baseline, while for the Istella dataset, simply selecting all features is substantially more effective than our feature selection. For future work, we believe that accurately determining the number of to-be-selected features is an important next step, as this would help to not reduce the effectiveness in the Istella scenario.

Acknowledgments

References

[1] T. Liu, Learning to Rank for Information Retrieval, Springer, 2011. URL: https://doi.org/10.1007/978-3-642-14267-3. doi:10.1007/978-3-642-14267-3.

[2] J. Lin, R. F. Nogueira, A. Yates, Pretrained Transformers for Text Ranking: BERT and Beyond, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2021. URL: https://doi.org/10.2200/S01123ED1V01Y202108HLT053. doi:10.2200/S01123ED1V01Y202108HLT053.

[3] D. Dato, S. MacAvaney, F. M. Nardini, R. Perego, N. Tonellotto, The istella22 dataset: Bridging traditional and neural learning to rank evaluation, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 3099–3107. URL: https://doi.org/10.1145/3477495.3531740. doi:10.1145/3477495.3531740.

[4] M. Fröbe, S. Günther, M. Probst, M. Potthast, M. Hagen, The Power of Anchor Text in the Neural Retrieval Era, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval. 44th European Conference on IR Research (ECIR 2022), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2022.

[5] A. Pasin, M. F. Dacrema, P. Cremonesi, N. Ferro, Quantumclef - quantum computing at CLEF, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V, volume 14612 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 482–489. URL: https://doi.org/10.1007/978-3-031-56069-9_66. doi:10.1007/978-3-031-56069-9_66.

[6] A. Pasin, M. Ferrari Dacrema, P. Cremonesi, N. Ferro, QuantumCLEF 2024: Overview of the Quantum Computing Challenge for Information Retrieval and Recommender Systems at CLEF, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, September 9th to 12th, 2024, 2024.

[7] A. Pasin, M. Ferrari Dacrema, P. Cremonesi, N. Ferro, Overview of QuantumCLEF 2024: The Quantum Computing Challenge for Information Retrieval and Recommender Systems at CLEF, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, 2024.

[8] M. F. Dacrema, A. Pasin, P. Cremonesi, N. Ferro, Quantum computing for information retrieval and recommender systems, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V, volume 14612 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 358–362. URL: https://doi.org/10.1007/978-3-031-56069-9_47. doi:10.1007/978-3-031-56069-9_47.

[9] B. Efron, R. Tibshirani, An Introduction to the Bootstrap, Springer, 1993. URL: https://doi.org/10.1007/978-1-4899-4541-9. doi:10.1007/978-1-4899-4541-9.

[10] M. Fröbe, L. Gienapp, M. Potthast, M. Hagen, Bootstrapped nDCG Estimation in the Presence of Unjudged Documents, in: Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), volume 13980 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2023, pp. 313–329. doi:10.1007/978-3-031-28244-7_20.

[11] C. Buckley, D. Dimmick, I. Soboroff, E. M. Voorhees, Bias and the limits of pooling for large collections, Inf. Retr. 10 (2007) 491–508. URL: https://doi.org/10.1007/s10791-007-9032-x. doi:10.1007/S10791-007-9032-X.

[12] E. M. Voorhees, N. Craswell, J. Lin, Too many relevants: Whither cranfield test collections?, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 2970–2980. URL: https://doi.org/10.1145/3477495.3531728. doi:10.1145/3477495.3531728.

[13] J. Zobel, How reliable are the results of large-scale information retrieval experiments?, in: W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, J. Zobel (Eds.), SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia, ACM, 1998, pp. 307–314. URL: https://doi.org/10.1145/290941.291014. doi:10.1145/290941.291014.

[14] J. Savoy, Statistical inference in retrieval effectiveness evaluation, Inf. Process. Manag. 33 (1997) 495–512. URL: https://doi.org/10.1016/S0306-4573(97)00027-7. doi:10.1016/S0306-4573(97)00027-7.

[15] M. D. Smucker, J. Allan, B. Carterette, A comparison of statistical significance tests for information retrieval evaluation, in: M. J. Silva, A. H. F. Laender, R. A. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, A. O. Falcão (Eds.), Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007, ACM, 2007, pp. 623–632. URL: https://doi.org/10.1145/1321440.1321528. doi:10.1145/1321440.1321528.

[16] T. Sakai, Evaluating evaluation metrics based on the bootstrap, in: E. N. Efthimiadis, S. T. Dumais, D. Hawking, K. Järvelin (Eds.), SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006, ACM, 2006, pp. 525–532. URL: https://doi.org/10.1145/1148170.1148261. doi:10.1145/1148170.1148261.

[17] T. Sakai, On the reliability of information retrieval metrics based on graded relevance, Inf. Process. Manag. 43 (2007) 531–548. URL: https://doi.org/10.1016/j.ipm.2006.07.020. doi:10.1016/J.IPM.2006.07.020.

[18] J. Zobel, L. Rashidi, Corpus bootstrapping for assessment of the properties of effectiveness measures, in: M. d'Aquin, S. Dietze, C. Hauff, E. Curry, P. Cudré-Mauroux (Eds.), CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, ACM, 2020, pp. 1933–1952. URL: https://doi.org/10.1145/3340531.3411998. doi:10.1145/3340531.3411998.

[19] G. V. Cormack, T. R. Lynam, Statistical precision of information retrieval evaluation, in: E. N. Efthimiadis, S. T. Dumais, D. Hawking, K. Järvelin (Eds.), SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006, ACM, 2006, pp. 533–540. URL: https://doi.org/10.1145/1148170.1148262. doi:10.1145/1148170.1148262.

[20] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182. URL: http://jmlr.org/papers/v3/guyon03a.html.

[21] M. B. Shirzad, M. R. Keyvanpour, A systematic study of feature selection methods for learning to rank algorithms, Int. J. Inf. Retr. Res. 8 (2018) 46–67. URL: https://doi.org/10.4018/IJIRR.2018070104. doi:10.4018/IJIRR.2018070104.

[22] A. Gigli, C. Lucchese, F. M. Nardini, R. Perego, Fast feature selection for learning to rank, in: B. Carterette, H. Fang, M. Lalmas, J. Nie (Eds.), Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016, Newark, DE, USA, September 12- 6, 2016, ACM, 2016, pp. 167–170. URL: https://doi.org/10.1145/2970398.2970433. doi:10.1145/2970398.2970433.

[23] M. F. Dacrema, F. Moroni, R. Nembrini, N. Ferro, G. Faggioli, P. Cremonesi, Towards feature selection for ranking and classification exploiting quantum annealers, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 2814–2824. URL: https://doi.org/10.1145/3477495.3531755. doi:10.1145/3477495.3531755.

[24] X. Geng, T. Liu, T. Qin, H. Li, Feature selection for ranking, in: W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, N. Kando (Eds.), SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007, ACM, 2007, pp. 407–414. URL: https://doi.org/10.1145/1277741.1277811. doi:10.1145/1277741.1277811.

[25] G. Hua, M. Zhang, Y. Liu, S. Ma, L. Ru, Hierarchical feature selection for ranking, in: M. Rappa, P. Jones, J. Freire, S. Chakrabarti (Eds.), Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010, ACM, 2010, pp. 1113–1114. URL: https://doi.org/10.1145/1772690.1772830. doi:10.1145/1772690.1772830.

[26] K. D. Naini, I. S. Altingövde, Exploiting result diversification methods for feature selection in learning to rank, in: M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky, K. Hofmann (Eds.), Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings, volume 8416 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 455–461. URL: https://doi.org/10.1007/978-3-319-06028-6_41. doi:10.1007/978-3-319-06028-6_41.

[27] A. Pasin, M. F. Dacrema, P. Cremonesi, N. Ferro, qclef: A proposal to evaluate quantum annealing for information retrieval and recommender systems, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings, volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 97–108. URL: https://doi.org/10.1007/978-3-031-42448-9_9. doi:10.1007/978-3-031-42448-9_9.

[28] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Advances in Neural Information Processing Systems 30 (2017).

[29] C. Macdonald, N. Tonellotto, Declarative experimentation in information retrieval using pyterrier, in: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, 2020, pp. 161–168.

[30] D. Alexander, M. Fröbe, G. Hendriksen, F. Schlatt, M. Hagen, D. H. ad Martin Potthast, A. P. de Vries, Team openwebsearch at clef 2024: Longeval, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, September 9th to 12th, 2024, CEUR Workshop Proceedings, 2024.