SINAI team at QuantumCLEF 2025: Quantum Feature Selection Based on Energy with D-Wave

Lucas Molino-Piñar[1,*,†], Jaime Collado-Montañez[1,*,†] and Arturo Montejo-Ráez[1,†]

[1]Department of Computer Science (University of Jaén), Campus Las Lagunillas, s/n, Jaén, 23071, Spain

Abstract

This paper presents the SINAI team's approach to Task 1 (Feature Selection) at QuantumCLEF 2025, focusing on the use of quantum annealing with D-Wave hardware to tackle feature selection for learning-to-rank models. We formulate the feature selection problem as a Quadratic Unconstrained Binary Optimization problem based on mutual information to balance feature relevance and redundancy. Candidate feature subsets are generated by the quantum annealer and then post-processed through normalization and energy projection strategies to obtain robust feature rankings. Evaluation on the MQ2007 LETOR dataset demonstrates the potential of quantum computing to support effective feature selection in information retrieval tasks, despite current hardware limitations. Our results highlight promising directions for integrating quantum optimization in practical machine learning workflows.

Keywords

Quantum Annealing, Feature Selection, QUBO, D-Wave, Mutual Information, Quantum Computing

1. Introduction

The emergence of Quantum Computing (QC) represents a significant shift in computational science, offering the potential to solve complex problems more efficiently by leveraging quantum-mechanical principles such as superposition and entanglement. Unlike classical systems based on binary bits, QC utilizes quantum bits (qubits) to explore vastly larger solution spaces, opening new opportunities for fields like Information Retrieval (IR) and Recommender Systems (RS). While QC is still constrained by hardware limitations, recent advances have made quantum devices increasingly accessible, enabling researchers to begin exploring practical applications.

To support this exploration, the QuantumCLEF lab, part of the CLEF conference series, investigates the applicability of quantum algorithms to IR and RS challenges. The lab focuses on Quantum Annealing (QA)—a quantum optimization method—and provides participants with access to D-Wave's quantum hardware [1], along with tools for developing and testing quantum solutions.

In QuantumCLEF 2025 [2, 3], participants address three core problems in the information access domain: Feature Selection, Instance Selection, and Clustering. Our team, SINAI, took part in Task 1: Feature Selection, which aims to identify the most relevant subset of features for training learning-to-rank models. This task is essential for improving model efficiency and performance by reducing input dimensionality.

Our approach focused on framing the feature selection problem as a Quadratic Unconstrained Binary Optimization (QUBO) problem, suitable for quantum annealers. We employed a mutual information-based formulation to balance feature relevance and redundancy, and used D-Wave's quantum hardware to generate candidate solutions. We also developed a suite of post-processing techniques to aggregate these solutions into stable feature rankings, which were then evaluated using LambdaMART models on benchmark datasets.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

lmolino@ujaen.es (L. Molino-Piñar); jcollado@ujaen.es (J. Collado-Montañez); amontejo@ujaen.es (A. Montejo-Ráez)

0009-0009-5990-9802 (L. Molino-Piñar); 0000-0002-9672-6740 (J. Collado-Montañez); 0000-0002-8643-2714 (A. Montejo-Ráez)

The remainder of this paper is organized as follows: Section 2, Methodology, describes our QUBO formulation and post-processing strategies. Section 3, Results, presents our experimental evaluation. Section 4, Conclusions, summarizes our contributions and outlines directions for future work.

2. Methodology

2.1. Task Description and Dataset

The main objective of this task is to select the most relevant features from the MQ2007 dataset, which belongs to the LETOR 4.0 collection [4], in order to train a model based on LambdaMART [5] and achieve the best possible performance in document ranking.

The MQ2007 dataset originates from the LETOR (LEarning TO Rank) benchmark, developed by Microsoft Research Asia, and is specifically designed for learning-to-rank tasks. MQ2007 uses documents from the Gov2 web collection, which comprises approximately 25 million pages, and queries derived from the Million Query track (TREC 2007) [6]. The dataset contains around 1700 queries, each associated with documents labeled with different levels of relevance.

Each data instance corresponds to a query-document pair, characterized by 46 features that include classical IR and NLP measures such as TF-IDF, BM25, language models (LMIR), PageRank, document length, and URL structure. The dataset is preprocessed and normalized at the query level (QueryLevelNorm version), which allows for its direct use in machine learning models.

2.2. Formulation of the QUBO Matrix

Quadratic Unconstrained Binary Optimization (QUBO) is a mathematical framework used to express a wide range of combinatorial optimization problems [7] [8] [9]. In its standard form, the goal is to find a binary vector x that minimizes a quadratic function defined by a matrix Q:

$$\min x^T Q x \tag{1}$$

where:

- $x \in \{0,1\}^n$ is a binary vector of size n,
- $Q \in \mathbb{R}^{n \times n}$ is a symmetric matrix of coefficients that encodes the interactions between variables.

This formulation enables the modeling of problems involving selection, assignment, or combination. It is especially suitable for solution via metaheuristic algorithms or specialized hardware such as quantum computing systems. To incorporate constraints (e.g., a maximum number of selected elements), penalty terms are added to the objective function.

In this work, we implemented a formulation based on Mutual Information, called MIQUBO, as provided in the baseline for the feature selection problem. This formulation reflects both the relevance and redundancy between pairs of features [10].

The diagonal terms Q_{ii} encode the individual relevance of each feature X_i with respect to the target variable Y:

$$Q_{ii} = -I(X_i; Y) = -(H(X_i) - H(X_i \mid Y)) \tag{2}$$

Here, $H(\cdot)$ denotes the entropy of a variable, and $H(\cdot \mid \cdot)$ is the conditional entropy.

The off-diagonal terms Q_{ij} represent the redundancy between pairs of features, penalizing the joint selection of two variables that provide similar information. These are defined using conditional mutual information:

$$Q_{ij} = -I(X_i; Y \mid X_j) = -(H(X_i \mid X_j) - H(X_i \mid X_j, Y)) \tag{3}$$

The negative sign is a necessary adaptation for the matrix to be used in quantum hardware, aiming to minimize energy.

To impose a constraint on the total number of selected features (e.g., selecting exactly k), a quadratic penalty term is incorporated:

$$\left(\sum_i x_i - k\right)^2 \tag{4}$$

Thus, the complete MIQUBO objective function is defined as:

$$\min \sum_i Q_{ii} x_i + \sum_{i<j} Q_{ij} x_i x_j + \lambda \left(\sum_i x_i - k\right)^2 \tag{5}$$

Here, λ is a regularization hyperparameter that controls the strength of the penalty for selecting more or fewer than k features, which typically takes high values.

2.3. Post-processing Based Approach

Once the problem was defined and the QUBO matrix was built based on mutual information and conditional mutual information, our approach focused on the postprocessing of the solutions provided by the quantum hardware.

The quantum hardware used, via the D-Wave library, returns a set of candidate solutions, each accompanied by its corresponding energy value, which is determined by the objective function the system aims to minimize. The number of solutions returned is controlled by a parameter called the number of reads. For example, if this is set to 100, the quantum system will return 100 solutions, each with its associated energy.

Our approach is to leverage this diversity of solutions to build a more robust and stable ranking of features, taking into account both the frequency with which each feature appears in the selected solutions and the energy associated with those solutions.

To analyze the behavior of the quantum system for different subset sizes of selected features, the program was executed for a selected range of k values, rather than uniformly across all possible values. These values were chosen strategically to explore the solution space broadly in an initial phase, with the objective of identifying regions that showed promising performance. Based on this preliminary analysis, subsequent executions focused on narrower intervals to refine the search around the most favorable configurations. Each solution set obtained corresponds to a partial feature selection for a specific value of k, and these were later used to build a global ranking that reflects the overall importance of each feature across the explored configurations.

To guide the selection of the optimal value of k, the minimum energy obtained in each of the runs was recorded and analyzed. These values form a discrete energy profile across the range of k. By interpolating or fitting a smooth curve to these minimum energy values, it becomes possible to identify trends and locate the global minimum more precisely. The value of k corresponding to this global minimum is then selected as the most suitable number of features to retain.

To achieve this aggregation, two postprocessing strategies were designed and implemented, aiming to transform energy values into normalized or projected importance scores that enable consistent feature ranking.

Below are the two implemented strategies:

2.3.1. Direct normalization

In this strategy, the energy values returned by the quantum system were linearly normalized to the interval $[0, 1]$. To do this, the minimum and maximum energy values among all solutions obtained across the 45 runs were identified, and the following transformation was applied:

$$E_{\text{norm}} = \frac{E - E_{\min}}{E_{\max} - E_{\min}} \tag{6}$$

where E is the energy of an individual solution.

Then, for each feature, the normalized energy values of the solutions in which it appears were summed and divided by the number of times the feature occurred, resulting in its average normalized energy score. Once all feature scores were obtained, a final normalization to the $[0, 1]$ range was applied to these average scores. Finally, the features were ranked in ascending order of normalized average energy, based on the assumption that lower energy implies higher relevance.

2.3.2. Signed energy projection

In this approach, a differentiated projection was applied based on the sign of the energy:

- **Negative energy values** were projected linearly to the interval $[-1, 0]$.
- **Positive energy values** were projected to the interval $[0, 1]$.

Let E be the energy value of a solution, and let $E_{\min}$ and $E_{\max}$ be the minimum negative energy and maximum positive energy observed, respectively. The projected energy E' was computed as:

$$E_{norm} = \begin{cases} \dfrac{E - E_{\min}}{E_{\min}}, & \text{if } E < 0 \\ \dfrac{E}{E_{\max}}, & \text{if } E \geq 0 \end{cases} \tag{7}$$

For each feature, the projected energy values of all the solutions in which it appears were summed and divided by the number of occurrences of that feature, resulting in its average projected energy score. This score was then linearly normalized to the range $[-1, 1]$, preserving the distinction between favorable (low-energy) and unfavorable (high-energy) contributions. Features were subsequently ranked in ascending order of their normalized scores, under the assumption that lower values indicate higher relevance.

2.4. Evaluation

The quality of the selected features will be evaluated by training a LambdaMART model and testing its performance on a held-out test set. The evaluation metric will be the normalized Discounted Cumulative Gain at rank 10 (nDCG@10), defined as:

$$\text{nDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}} \tag{8}$$

where DCG@k is the discounted cumulative gain at position k, and IDCG@k is the ideal discounted cumulative gain at position k, which represents the maximum possible DCG@k value [11].

3. Results

The generation of solutions was carried out using quantum computing techniques based on the quantum annealing paradigm. However, the process was constrained by technical limitations due to restricted access to the quantum hardware. These restrictions prevented us from exploring all the predefined possibilities; nonetheless, we present the analysis of our execution strategy.

Several executions were performed, each consisting of 100 reads, yielding 100 candidate solutions per run. The parameter of interest in this experiment was the number of selected features, denoted as K, which was varied in increments of 5 units, from $K = 5$ to $K = 45$. For each K, the energies of the 100 obtained solutions were collected.

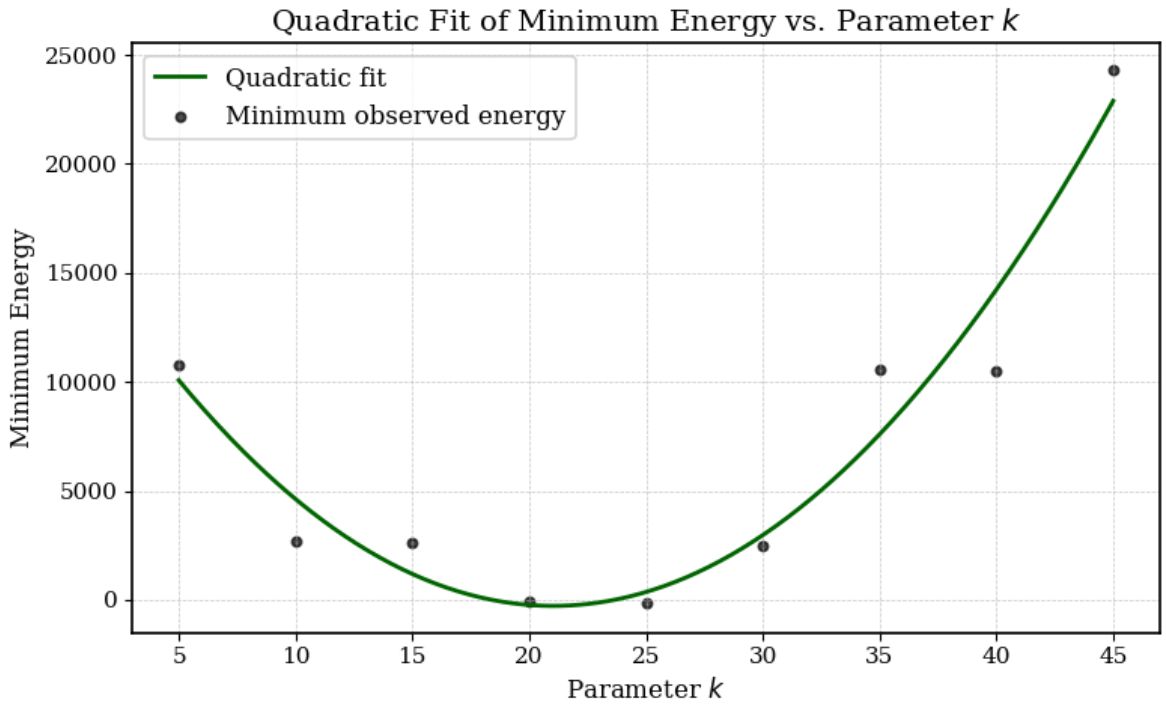

Figure 1: Quadratic fit of the minimum energy as a function of the parameter k. The energy values corresponding to different k values from 5 to 45 (increments of 5) were processed. For each k, the minimum observed energy was calculated. Then, a second-degree polynomial (quadratic) fit was performed on these minimum values to model the trend. The graph shows the minimum observed energy points (in black) and the fitted curve (in dark green).

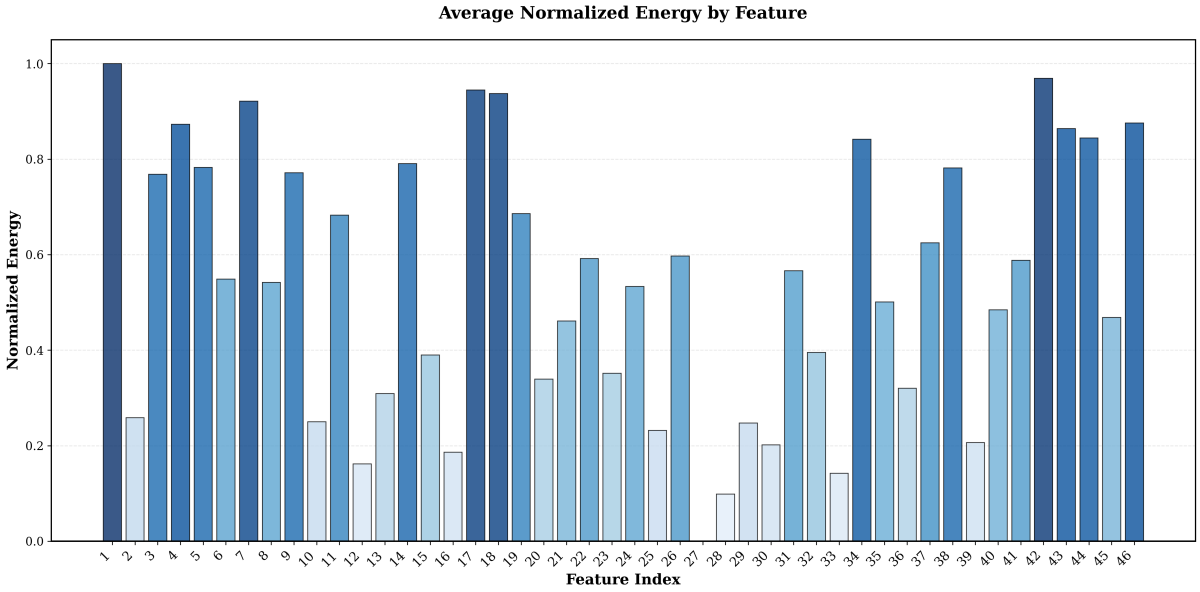

Figure 2: Normalized mean energy per feature, obtained from an initial energy metric normalized to the range [0, 1]. The original energy values were individually normalized and then aggregated per feature by averaging over all subsets where the feature appears. A final normalization was applied to the averages to rescale them again between 0 and 1. The features are numerically ordered according to their index. The bar colors correspond to a blue scale indicating the magnitude of the normalized energy, with darker blue representing higher values.

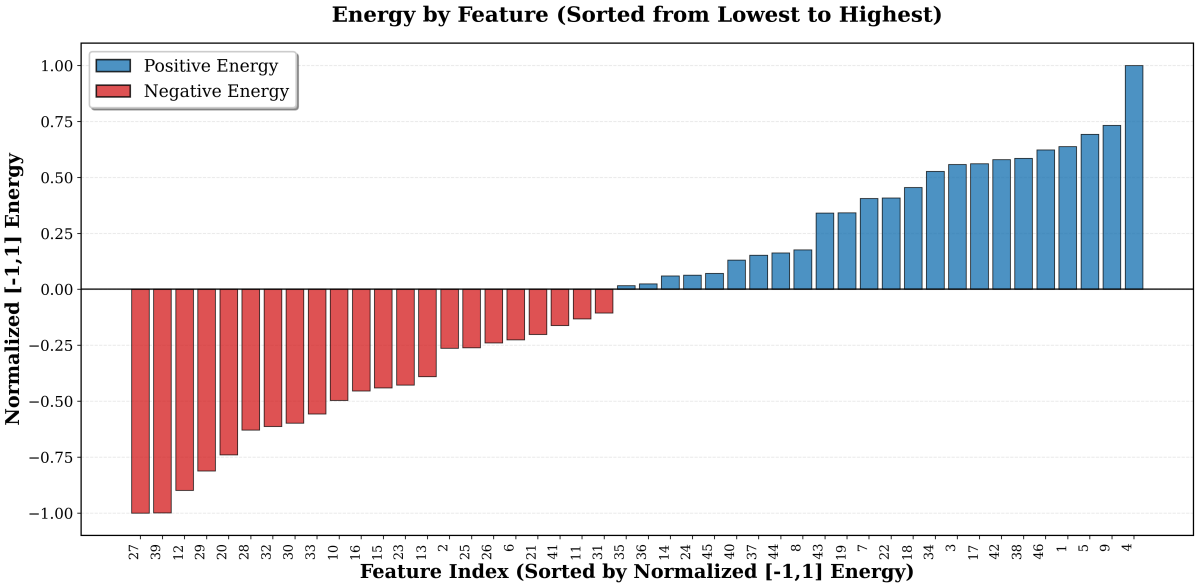

Figure 3: Distribution of normalized energy $[-1, 1]$ associated with each feature, sorted from lowest to highest. The energy was initially calculated from a global metric per feature subset and then distributed among individual features according to their frequency. A two-step normalization using a scaler was applied: first by sign (positive or negative) to preserve proportionality, and then scaling the final set to the range $[-1, 1]$. Red bars indicate features with negative energy, while blue bars represent those with positive energy.

After gathering all results, the relationship between the number of selected features and the minimum observed quantum energy was analyzed. This relationship was modeled by fitting a quadratic function. Figure 1 shows the fitted curve, illustrating how the energy evolves as a function of K. The minimum of this fitted curve indicates the value of K associated with the most energetically favorable configurations. This minimum was calculated analytically using the vertex formula for a parabola:

$$K^* = -\frac{b}{2a} \tag{9}$$

where a and b are the coefficients of the fitted quadratic function. The optimal number of features was found to be $K^* = 21.026$.

Following this, two post-processing strategies described in Section 2.3 were applied to select the final feature sets. The first strategy normalized the quantum solutions to a [0,1] range, generating a scoring vector for each feature, as shown in Figure 2.

Subsequently, the second post-processing strategy was applied, which used an alternative normalization method also described in the methodology. This technique enabled the generation of an alternative final ranking based on more robust statistical criteria. The computed scores revealed that there are 23 features with negative energy values after normalization. These scores were then linearly scaled to the range $[-1, 1]$, preserving the distinction between favorable (low-energy) and unfavorable (high-energy) contributions. The final ranking obtained from this normalization process is presented in Figure 3.

Although both generated rankings could not be directly evaluated due to technical and time limitations related to the execution environment, they were used as a reference to select the best solutions obtained via the quantum hardware. For this purpose, additional runs were performed in steps of 2 from $k = 21$ to $k = 29$ with 100 reads, as previously done. The same experiments were also run using quantum simulation with 3000 reads to compare the efficiency of each configuration.

The results shown in Table 1 reveal that the best performance, according to the ndcg@10 metric, was achieved with a quantum execution using $K = 21$ features and 100 reads, obtaining a score of 0.4580. This value aligns closely with the optimal feature count estimated from the quadratic fitting analysis.

In general, quantum executions with fewer reads tend to slightly outperform their simulated counterparts with a larger number of reads, suggesting that the actual quantum annealing process may

Table 1
Results sorted by K for different configurations of execution type and number of reads. The best `ndcg@10` result is highlighted in **bold**.

Type	Features (K)	Reads	ndcg@10
Simulated	29	3000	0.4491
Quantum	29	100	0.4528
Simulated	27	3000	0.4438
Quantum	27	100	0.4425
Simulated	25	3000	0.4510
Quantum	25	100	0.4550
Simulated	23	3000	0.4478
Quantum	23	100	0.4437
Simulated	21	3000	0.4530
Quantum	**21**	**100**	**0.4580**

capture more favorable solutions under certain configurations. Nonetheless, the differences across configurations are relatively small, indicating that the selected subset sizes ($K \in [21, 29]$) all yield competitive results.

4. Conclusions

This work presented the SINAI team's approach to the Feature Selection task at QuantumCLEF 2025, leveraging D-Wave's quantum annealing hardware to optimize feature selection for learning-to-rank models. By formulating the problem as a QUBO model based on mutual information, we effectively captured the trade-off between feature relevance and redundancy. Our post-processing strategies—direct normalization and signed energy projection—enabled us to derive stable and interpretable feature rankings from the quantum solutions.

The results, evaluated on the MQ2007 dataset using LambdaMART and nDCG@10 as the performance metric, demonstrated that quantum executions can yield competitive results, even under limited access constraints. The optimal subset of 21 features, as identified through a quadratic energy fitting process, achieved the highest score among tested configurations, suggesting that the energy landscape provided by quantum annealing correlates meaningfully with model performance.

Although hardware and execution limitations prevented exhaustive experimentation, our findings indicate that quantum annealing holds promise as a tool for feature selection in information retrieval contexts. Future work will focus on expanding the evaluation, incorporating other datasets, and exploring hybrid quantum-classical methods to further exploit the advantages of quantum optimization.

Acknowledgments

This work has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, and by the scholarship (FPI-PRE2022-105603) from the Ministry of Science, Innovation and Universities of the Spanish Government. Also, this work has been funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Pasin, N. Ferro, Kimera: From evaluation-as-a-service to evaluation-in-the-cloud, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padova, Italy, July 13-18, 2025, ACM, 2025. URL: https://doi.org/10.1145/3726302.3730298. doi:10.1145/3726302.3730298.

[2] A. Pasin, M. F. Dacrema, W. Cuhna, M. A. Gonçalves, P. Cremonesi, N. Ferro, Quantumclef 2025: Overview of the second quantum computing challenge for information retrieval and recommender systems at CLEF, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[3] A. Pasin, M. F. Dacrema, W. Cuhna, M. A. Gonçalves, P. Cremonesi, N. Ferro, Overview of quantumclef 2025: The second quantum computing challenge for information retrieval and recommender systems at CLEF, in: J. Carrillo-de-Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, 2025.

[4] T. Qin, T.-Y. Liu, Introducing letor 4.0 datasets, 2013. URL: https://arxiv.org/abs/1306.2597. arXiv:1306.2597.

[5] C. J. Burges, From RankNet to LambdaRank to LambdaMART: An Overview, Technical Report MSR-TR-2010-82, 2010. URL: https://www.microsoft.com/en-us/research/publication/from-ranknet-to-lambdarank-to-lambdamart-an-overview/.

[6] J. Allan, B. Carterette, B. Dachev, J. Aslam, V. Pavlu, E. Kanoulas, Million query track 2007 overview, 2007.

[7] A. Lucas, Ising formulations of many np problems, Frontiers in Physics 2 (2014). URL: http://dx.doi.org/10.3389/fphy.2014.00005. doi:10.3389/fphy.2014.00005.

[8] F. Glover, G. Kochenberger, Y. Du, A tutorial on formulating and using qubo models, 2019. URL: https://arxiv.org/abs/1811.11538. arXiv:1811.11538.

[9] C. Mcgeoch, Adiabatic quantum computation and quantum annealing: Theory and practice, Synthesis Lectures on Quantum Computing 5 (2014) 1–93. doi:10.2200/S00585ED1V01Y201407QMC008.

[10] Z. Zhang, E. Hancock, Mutual information criteria for feature selection, volume 7005, 2011. doi:10.1007/978-3-642-24471-1_17.

[11] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Trans. Inf. Syst. 20 (2002) 422–446. URL: https://doi.org/10.1145/582415.582418. doi:10.1145/582415.582418.