Team qIIMAS on Task 2 - Clustering

Notebook for the QuantumCLEF Lab at CLEF 2024

William Alvarez-Giron[1], Jorge Téllez-Torres[1,2], Javier Tovar-Cortes[1,2] and
Helena Gómez-Adorno[1]

[1] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Ciudad Universitaria, 04510, CDMX, México.

[2] Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad Universitaria, 04510, CDMX, México.

Abstract

This paper describes our participation in Task 2 of the QuantumCLEF lab, where we explored the use of quantum annealing for clustering 6486 document embeddings. We employed a Quadratic Unconstrained Binary Optimization (QUBO) formulation to address the NP-hard problem of k-medoids clustering, optimizing the selection of cluster centers without explicitly assigning each data point to a cluster. Our approach aimed to minimize pairwise dissimilarities among selected medoids while maximizing their separation, ensuring well-defined and diverse clusters. To manage the limitations of current quantum hardware, we implemented a hierarchical clustering procedure: initially reducing the dataset using classical methods and then applying quantum annealing for final medoid optimization. The results indicate that a combination of simulated annealing for initial clustering, followed by quantum annealing, yielded the best outcomes, achieving a Davies-Bouldin index of 4.6798 and an nDCG@10 of 0.5349 when using $k = 50$ clusters. This hierarchical approach underscores the potential of quantum annealing in enhancing clustering quality despite current hardware constraints. Our findings suggest that quantum-enhanced clustering methods could significantly advance the field of information retrieval and recommender systems by providing more efficient and accurate data organization techniques.

Keywords

Quantum Annealing, Clustering, Information Retrieval, Recommender Systems

1. Introduction

Clustering is an unsupervised machine learning technique that involves grouping similar data points together [1]. It has many applications in information retrieval, such as organizing large document collections, providing similar search results to a query, and dividing users according to interests to build user models [2]. However, clustering can be a complex and computationally expensive task, especially when dealing with large datasets [3, 4].

Quantum Annealing (QA) is an emerging technology that leverages quantum mechanical effects to find optimal solutions to certain optimization problems [5, 6]. Several research studies have explored how quantum annealing can be applied to challenging clustering problems [7, 8, 9, 10]. Quantum algorithms have the potential to provide better quality clusterings in less time compared to classical clustering algorithms [8, 11, 12].

This paper discusses our participation in Task 2 of the *QuantumCLEF (qCLEF)* lab, an innovative evaluation platform designed to explore the potential of QA in the domains of *Information Retrieval (IR)* and *Recommender Systems (RS)*. The goal of this task is to use quantum annealing to group document embeddings to facilitate the browsing of large collections [13, 14, 15]. We aim to evaluate the effectiveness and efficiency of a quantum annealing approach compared to a classical clustering baseline.

For our QA method, we formulate the clustering problem as a *Quadratic Unconstrained Binary Optimization (QUBO)* that can be solved by a quantum annealer. We tested our approach on the provided datasets of sentence embeddings from ANTIQUE [14, 15]. The quality of the obtained clusters was evaluated using the Davies-Bouldin index to measure the overall quality of the cluster. Additionally,

CLEF 2024: Conference and Labs of the Evaluation Forum, September 9–12, 2024, Grenoble, France

wikkilicht@ciencias.unam.mx (W. Alvarez-Giron); jorge97@ciencias.unam.mx (J. Téllez-Torres); javtovar@ciencias.unam.mx (J. Tovar-Cortes); helena.gomez@iimas.unam.mx (H. Gómez-Adorno)

we use the clusterings to retrieve the most relevant documents for a set of test queries and measure retrieval performance using nDCG@10.

Our focus within this framework has been on leveraging QA to tackle the computationally intensive task of document clustering. This process, essential for organizing large datasets and improving user interaction through more precise recommendation systems, poses significant challenges due to the sheer volume and complexity of the data. Traditional clustering methods, while effective to some extent, often struggle under the weight of modern data demands, necessitating a quantum approach that could potentially revolutionize our capabilities. The qCLEF lab provided a unique opportunity to compare the performance of QA-based clustering algorithms against their classical counterparts (*Simulated Annealing (SA)*), using both the quantum resources of D-Wave machines accessing by the CINECA platform and traditional computational methods.

The paper is organized as follows. Section 2 introduces related works; Section 3 describes our approach; Section 4 explains our experimental setup; Section 5 discusses our main findings; finally, Section 6 draws some conclusions and outlooks for future work.

2. Related Work

QA, an optimization algorithm proposed by Kadowaki and Nishimori in 1998 [5], leverages the quantum tunneling effect to solve complex optimization problems more efficiently than classical methods and has gained significant attention in recent years. The D-Wave machine, a commercial quantum annealer, has been extensively used to implement quantum annealing algorithms in various domains, including clustering, IR, and RS [16, 17].

To utilize QA, the optimization problem must first be formulated as a QUBO problem. This formulation allows the problem to be mapped onto the quantum annealer's architecture. Once the problem is expressed as a QUBO, QA operates by encoding the solution into the ground state of a quantum system. The process involves initializing the system into a superposition of all possible states and gradually evolving it towards the ground state using a technique known as adiabatic evolution. This method has shown promise in solving combinatorial optimization problems, where traditional algorithms struggle due to computational complexity [18].

In the realm of optimization, QA has shown significant potential. For example, it has been applied to solve complex scheduling problems in aviation and manufacturing, optimizing resource allocation and process planning [17, 19]. Clustering, a crucial technique in unsupervised learning, involves grouping similar objects. To solve clustering problems using QA on D-Wave machines, the task must first be formulated as a QUBO problem. This formulation enables the problem to be efficiently mapped onto the quantum annealer's architecture. Recent studies have extended QA's applicability to graph clustering, enhancing traditional methods by reducing computational time while maintaining or improving accuracy [20].

In the context of clustering, QA has been applied to optimize the clustering objective function. Kurihara et al. [7] introduced a quantum annealing-based approach for clustering high-dimensional data, demonstrating its superiority over classical clustering algorithms. Arthur et al. [10] proposed a balanced k-means clustering algorithm implemented on an adiabatic quantum computer, showcasing its ability to handle large-scale datasets efficiently.

In general, *Quantum Computing (QC)* has emerged as a promising approach to improve IR and RS [21]. For example, Chakrabarty et al. [22] extended the Grover search algorithm to approximate algorithms, introducing the Dynamic Grover search algorithm to define goals in recommendation systems and optimize various problems, providing a quadratic speedup over traditional classical approaches. Bhagawati [23] proposed a quantum-inspired document ranking algorithm using feature selection methodology, leveraging the unique properties of quantum systems. Particularly, QA has also gained attention for its potential to address challenges in IR and RS. Ferrari Dacrema et al. [24] utilized QA to select characteristics in ranking and classification tasks, demonstrating its potential to improve the performance of machine learning models in IR. Nembrini et al. [25] proposed a quantum annealing approach for feature selection in RS. The feature selection problem was formulated as a QUBO problem, which is well suited to solve on the D-Wave machine. Finally, Ferrari Dacrema et al. [26] proposed a formulation of the carousel selection problem for black box recommenders that can be effectively solved using a D-Wave quantum annealer. Using the adiabatic quantum computing paradigm and the ability of the D-Wave machine to solve NP-hard optimization problems, the authors demonstrated the potential of quantum computing in optimizing the selection of recommendation carousels while accounting for the interaction between different recommendation lists and facilitating user exploration of the catalog.

Despite promising results, QA still faces challenges due to the limitations of current quantum hardware. Ongoing research focuses on developing more efficient quantum annealing algorithms and addressing the scalability issues of quantum machines [16, 17, 27]. As quantum technology advances, it is expected that QA will play an increasingly important role in clustering applications in various domains.

To sum up, QA has emerged as a powerful tool for clustering and optimization, with successful applications in IR and RS. The D-Wave machine has been instrumental in implementing QA algorithms and exploring their potential. However, more research is necessary to fully harness the capabilities of QA and overcome the limitations of current quantum hardware.

3. Methodology

For quantum annealing-based clustering, we follow the approach of Bauckhage et al. [9], which addresses the problem of identifying k medoids among n data points without explicitly assigning each data point to a cluster. This method leverages the capabilities of quantum annealing to optimize the selection process, thus enhancing computational efficiency in solving the problem k-medoids, recognized as NP-hard.

The essence of the approach is captured in the Quadratic Unconstrained Binary Optimization (QUBO) formulation, designed to be executed on a quantum annealer. The formulation abstracts the clustering problem into finding an optimal set of cluster centers (medoids) by minimizing an objective function that represents the clustering criteria without directly clustering each data point.

The QUBO for the k-medoids problem is formulated as:

$$z^* = \text{argmin}_{z \in \{0,1\}^n} \left(\gamma \mathbf{z}^T \mathbf{1} \mathbf{1}^T \mathbf{z} - \frac{\alpha}{2} \mathbf{z}^T \boldsymbol{\Delta} \mathbf{z} + \beta \mathbf{z}^T \boldsymbol{\Delta} \mathbf{1} - 2\gamma k \mathbf{1} \right) \tag{1}$$

where $\boldsymbol{\Delta}$ is a matrix where each element Δ_{ij} quantifies the dissimilarity between data points i and j, based on the pairwise distances between the data points. This QUBO formulation seeks to minimize the dissimilarity among the selected medoids while ensuring that the selected points are spread across the entire dataset.

The function embedded within the QUBO aims to balance two critical objectives:

1. Minimizing the sum of pairwise dissimilarities among the chosen medoids, thereby ensuring that the clusters are tight and well-defined, weighted by the variable β.
2. Maximizing the separation between the selected medoids to enhance the diversity and coverage of the clusters in the data set, with the contribution weighted by the variable α.

A crucial constraint in this QUBO formulation is to ensure that exactly k medoids are selected from the set of n points. This is implemented via the constraint:

$$\sum_{i=1}^{n} z_i = k \tag{2}$$

This condition is essential for preserving the accuracy of the clustering procedure, guaranteeing that precisely k clusters are created, each represented by a single medoid. In our QUBO, the contribution to this constraint is represented by γ. The parameter γ acts as a Lagrange multiplier, which is crucial in balancing the minimization of the objective function against the constraint that the number of selected medoids is equal to k.

The settings of γ, along with α and β, help fine-tune the performance of the clustering algorithm on quantum annealing hardware. In our experiments, we fix the values of the hyperparameters $\alpha = 1/k$, $\beta = 1/n$, and $\gamma = 2$ in the QUBO formulation. These values are chosen according to the guidelines provided by Bauckhage et al. [9] to ensure a balanced contribution of the different terms in the objective function. The resulting QUBO is then submitted to a D-Wave quantum annealer for optimization. The binary vector z^* obtained from the quantum annealer indicates the selected medoids, with $z_i = 1$ denoting that the data point i is a medoid. This sophisticated approach provided by Bauckhage et al. not only simplifies the clustering process but also harnesses the power of quantum computing to address scalability and efficiency, presenting a significant advancement in clustering large and complex datasets.

Before applying the QUBO formulation on the quantum annealer, we experimented with different sizes of prior clusters to reduce the number of data points. This hierarchical approach allows us to handle larger datasets that exceed the capacity of the quantum hardware. We explore the impact of size reduction on clustering performance and solution quality.

The hierarchical clustering procedure, illustrated in Figure 1, works as follows:

1. We start with the original dataset of n sentence embeddings.
2. We apply a classical clustering algorithm (e.g., k-means, pairing-point method or simulated annealing) to group embeddings into m clusters, where $m < n$.
3. We use the m representative vectors as input to the QUBO formulation and solve it using the D-Wave quantum annealer.
4. The resulting medoids obtained from the quantum annealer are then used to assign each of the original n data points to its nearest medoid, forming the final clusters.

We vary the size m of the prior clustering to investigate the trade-off between the reduction of problem size and the quality of the clustering. Smaller values of m lead to a more compact representation of the dataset but may lose some fine-grained details. Larger values of m preserve more information, but increase computational complexity and may exceed the capacity of the quantum annealer.

By comparing the clustering results obtained with different sizes of prior clusterings, we aim to find a balance between the efficiency of the quantum approach and the effectiveness of the resulting clusters. We evaluated the quality of the final clusters using the Davies-Bouldin index and the nDCG@10 metric, as described in the evaluation measures section.

This hierarchical approach allows us to leverage the power of quantum annealing for the clustering task while addressing the limitations of current quantum hardware. By systematically exploring the impact of problem size reduction, we provide insights into the scalability and practicality of using quantum annealing for clustering in the context of information retrieval.

4. Experimental Setup

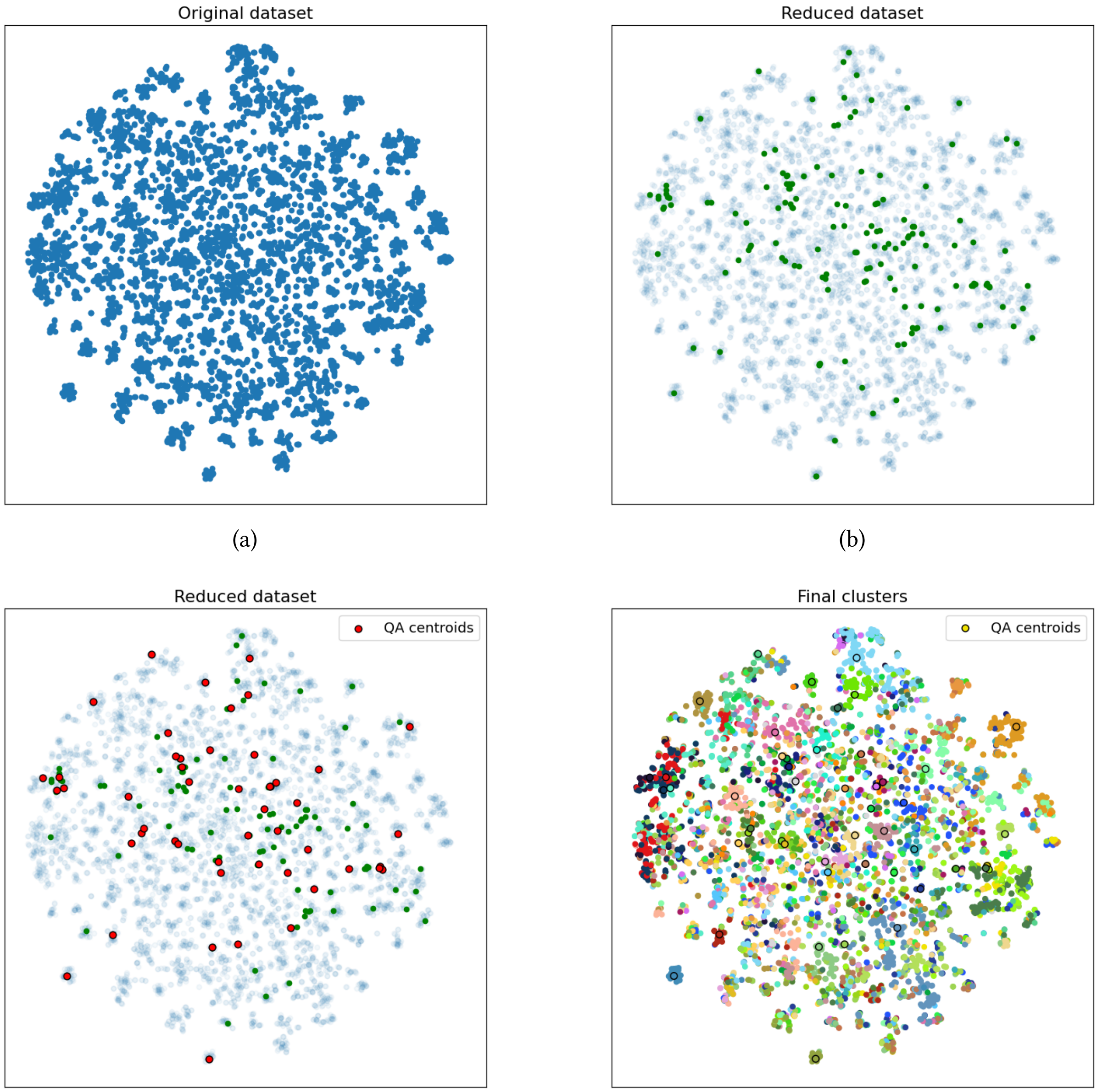
The clustering task in qCLEF focuses on using QA to group document embeddings to facilitate browsing of large collections. Specifically, the task involves obtaining a list of representative medoids for a given dataset of sentence embeddings.

4.1. Datasets

For this task, we used datasets derived from the ANTIQUE collection. The organizers provide two datasets [14, 15]:

- A larger dataset of 6486 sentence embeddings for clustering.
- A smaller dataset of 1651 sentence embeddings of training queries was used to measure the generalization of the clustering methods during experimentation.

The sentence embeddings provided by the organizers were generated using a transformer model to convert each sentence into a dense vector representation.



(a)

(b)

(c)

(d)

Figure 1: The figure illustrates the hierarchical clustering procedure: (a) The original dataset is visualized using t-SNE projection, showing the distribution of data points. (b) The dataset is reduced, highlighting the 150 medoids obtained by applying the k-medoids algorithm. (c) QA is used to obtain the centroids within the reduced dataset, marked in red. (d) The final clusters are displayed, with QA centroids highlighted.

4.2. Evaluation Metrics

Two main metrics are used to evaluate the quality of the clusters:

1. Davies-Bouldin Index: This is an internal evaluation metric utilized to assess the quality of clustering. It quantifies the average similarity ratio of each cluster to its most similar counterpart, considering both the within-cluster scatter (compactness) and the between-cluster separation. A lower DBI value indicates better clustering performance, characterized by compact and well-separated clusters.
2. Normalized Discounted Cumulative Gain (nDCG@10): This is an external evaluation metric that assesses the usefulness of the clustering for a retrieval task. For each query, the most relevant cluster is identified based on the similarity between the query embedding and cluster centroids.

The documents belonging to that cluster are then ranked based on their similarity to the query. So, the nDCG@10 metric measures the quality of the top 10 retrieved documents in terms of relevance. These relevance evaluations were calculated by the organizers after submitting the clustering results from the lab. A higher nDCG signifies that the clustering is more beneficial for retrieval tasks [14, 15].

4.3. Procedure

To apply quantum annealing to the clustering task, we formulate it as a QUBO problem that can be solved with a quantum annealer. We adopt a similar approach to that proposed by Bauckhage et al. [9]. In our QUBO formulation, we define binary variables z_i that indicate whether a data point i is selected as a medoid.

One challenge of the QUBO formulation is that it requires the specification of the number of clusters (k) in advance. To address this, participants are encouraged to experiment with $k = 10, 25$ or 50 values and use a cluster validity index such as Davies-Bouldin to select the best number of clusters. The organizers provide tutorials and code templates to assist participants in implementing the QUBO formulation and interfacing with the D-Wave quantum annealer. Participants have the flexibility to explore alternative problem formulations and strategies for mapping the clustering task to the quantum hardware, taking into account the specific characteristics and limitations of the quantum annealer.

To simulate the quantum annealing process on a classical computer, we use the SA algorithm, which can be implemented using the SimulatedAnnealingSampler method from the D-Wave sampler library. This allows us to test and debug our QUBO formulation locally before running on the actual quantum hardware.

For our quantum experiments, we utilized the D-Wave Advantage quantum annealer, which was made available through CINECA. In this lab, we were allocated 120 s of *Quantum Processing Unit (QPU)*. The device comprises over 5000 qubits organized in a Pegasus topology, where each qubit is linked to 15 others. To interact with the quantum annealer, we employed the dwave-system library, particularly the DWaveSampler and EmbeddingComposite classes. Despite this advanced architecture, the current hardware limitations still present challenges in embedding large problems directly onto the quantum hardware. Consequently, only a few hundred variables can be embedded in the machine. One potential solution could be the dimensionality reduction of the document vectors, which are part of Task 1 of the qCLEF lab. To explore other approaches, we proposed a hierarchical clustering strategy. This approach entails initial clustering using classical algorithms to decrease the number of vectors and select representatives until the problem size is small enough to fit into the quantum hardware. One example is the LeapHybridSampler, which is already part of the Dwave library. Additionally, we considered other classical algorithms, including k-medoids, simulated annealing, and an alternative involving pairwise clustering iterations.

Our experimental procedure is as follows:

1. We first run the classical k-medoids baseline on the set to establish a reference performance level. We experiment with $k = 10, 25, 50$ (number of clusters) according to the guidelines.
2. We then implement our proposed QUBO formulation of the clustering problem and validate it using the SimulatedAnnealingSampler.
3. We run experiments using the D-Wave quantum annealer, comparing the direct embedding approach for small k and the hierarchical approach for larger k. We tune hyperparameters such as the annealing time and the number of reads.
4. We compare the performance of the classical and quantum approaches using the Davies-Bouldin index.
5. We analyze the results and discuss the benefits and limitations of the quantum annealing approach for the clustering task.

4.4. Code Repository

Our code is available in the following BitBucket repository: https://bitbucket.org/eval-labs/qc24-qiimas/src/main/. The repository is organized as follows:

- **antique_doc_embeddings.csv** : This file contains a subset of the ANTIQUE dataset. Includes sentences from Yahoo! Answers that have been converted into embeddings using a transformer model. There are 6486 vectors for the clustering in total.
- **antique_train_queries.csv** : This file also includes a portion of the ANTIQUE dataset. It contains 50 unique query embeddings along with their manually assigned relevance annotations to documents from the dataset.
- **annealings.py** : This module contains all the functions used in our Quantum and Simulated Annealing codes. As well as in our Hybrid Algorithm.
- **clustering-two-algorithms.ipynb** : This notebook presents two algorithms.The first one involves executing K-medoids clustering followed by Quantum Annealing on the derived medoids. In the second approach, we executed Classical Annealing prior to implementing Quantum Annealing on the obtained medoids.
- **clustering.ipynb** : This notebook performs clustering using Simulated or Hybrid Annealing, utilizing reduced data obtained through point-pairing clustering.
- **hybrid_algorithm.py** : In this notebook, we split the data into one-thirty-second of its original size using a function named 'half_points', which clusters through point-pairing. Subsequently, we performed Hybrid Annealing using the LeapHybridSampler function on the downsized data.
- **submissions/** : This directory holds .txt files with the cluster medoids derived from each algorithm. We assessed clusters with $k = 15, 25$, and 50.
- **query_results/** : This folder contains the query results and the nDCG@10 metric evaluated in our submissions, which have been provided to us by the organizers.

5. Results

Our findings, summarized in Table 1, provide valuable insights into the performance of various clustering approaches, particularly the hierarchical methods combining classical and quantum techniques. In particular, for the **hybrid annealing (hybrid)** method provided by the D-Wave machine, it should be emphasized that because of the restricted computational time available in the laboratory, we considered the minimum time required to execute the hybrid method based on the number of vectors. Within these constraints, the hybrid method included with D-Wave, despite its built-in optimization, did not outperform the classical **simulated annealing (simulated)** approach. This suggests that the current state of the hybrid quantum-classical algorithm in D-Wave may require further refinements to fully leverage the potential advantages of quantum annealing when dealing with limited computational resources.

Among the **hierarchical methods (+)** explored, the **point-pair clustering (pp)** for the initial clustering approach stands out for its simplicity and ease of implementation. By initially reducing the size of the dataset using point-pair clustering and subsequently applying quantum annealing, this method offers a straightforward way to handle larger datasets while harnessing the power of quantum computing. However, our experiments reveal that the point-pair clustering approach does not yield the best results in terms of clustering quality. In contrast, employing **k-medoids** for the initial clustering step prior to quantum annealing leads to superior performance. This finding highlights the importance of selecting an appropriate classical clustering algorithm to effectively reduce the problem size and provide a solid foundation for the quantum annealing process.

Notably, the best results were obtained by combining classical simulated annealing for the initial clustering followed by **quantum annealing (quantum)** for further optimization. This two-step approach consistently outperformed other methods across different values of k (number of clusters). The simulated+quantum method achieved the lowest Davies-Bouldin index of 4.6798 and the highest nDCG@10 score of 0.5349 when using 50 clusters. These results underscore the synergistic effect of leveraging both classical and quantum techniques in a hierarchical manner. The initial simulated annealing step effectively reduces the problem complexity, allowing the quantum annealing process to refine and optimize the clustering solution.

Table 1

Comparison of the models. The evaluated models include simulated, hybrid, k-medoids, quantum, pairing-point, and combinations of these. The classical counterparts of the models are shaded in gray. The parameters of the two models used in the hierarchical procedure are separated by a comma. DS indicates the size of the dataset used for clustering. TL specifies the time limit for computation in seconds for the hybrid annealing. NR denotes the number of states to read in the quantum and simulated annealing. M indicates the metric, and RS indicates the random state for the kmedoids algorithm. D indicates how many divisions in half were applied to the dataset using point-pair clustering. k denotes the number of clusters used in the model. The best results for each metric are highlighted. The hybrid model with a dataset size of 6486, a computation time limit of 15s, and 10 clusters achieves the best NDCG@10 score of 0.5682. The simulated+quantum model with a dataset size of 150 and 50 clusters achieves the best DB index of 4.6798. The baseline kmedoids results are at the bottom of the table. The k-medoids baseline and the pp+simulated model were used as references and were not submitted like the other models.

Model	Parameters	DB index	NDGC@10
kmedoids baseline	DS:6486 RS:42 M:Euclidian k=10	21.801	-
kmedoids baseline	DS:6486 RS:42 M:Euclidian k=25	16.6938	-
kmedoids baseline	DS:6486 RS:42 M:Euclidian k=50	17.5655	-
simulated	DS:6486 NR:200 k=10	6.6847	0.5622
hybrid	**DS:6486 TL:15s k=10**	**6.3121**	**0.5682**
simulated	DS:6486 NR:200 k=25	5.3368	0.546
hybrid	DS:6486 TL:15s k= 25	5.3509	0.5490
simulated	DS:6486 NR:200 k=50	4.7868	0.5068
hybrid	DS:6486 TL:15s k=50	4.8032	0.5564
hybrid	DS: 6486 TL:30s k=50	4.9537	0.5274
kmedoids+simulated	DS:6486 RS:42 M:Euclidian k=150 , DS:150 NR:200 k=50	5.4159	0.5065
kmedoids+quantum	DS:6486 RS:42 M:Euclidian k=150 , DS:150 NR:200 k=50	5.1842	0.5180
simulated+simulated	DS:6486 NR:200 k=150 , DS:150 NR:200 k=50	4.6978	0.5310
simulated+quantum	**DS:6486 NR:200 k=150 , DS:150 NR:200 k=50**	**4.6798**	**0.5349**
pp+simulated	DS:6486 D:5 , DS:203 NR:200 k=50	5.4606	-
pp+hybrid	DS:6486 D:5 , DS:203 TL:4s k=50	5.0868	0.5011

6. Conclusions and Future Work

This paper explores the use of quantum annealing for clustering document embeddings, as part of Task 2 in the qCLEF lab at CLEF 2024. To address the limitations of current quantum hardware and the restricted QPU time available in the laboratory setting, we proposed a hierarchical approach that combines classical pre-processing with quantum annealing, showing promising results. The performance of simulated annealing models was robust, as reflected in the Davies-Bouldin (DB) index. However, models combining simulated and quantum annealing slightly surpassed the purely classical approach in cluster quality. These findings highlight the potential of quantum annealing in enhancing cluster quality.

Our results highlight the advantages of a hierarchical approach, which addresses the limitations of current quantum technology by first reducing the problem size using classical methods before applying quantum annealing. This approach not only optimizes the use of scarce quantum resources, but also facilitates the handling of larger datasets. Future research should focus on improving these techniques by developing more advanced pre-processing methods to enhance the quantum annealing process. Exploring advanced dimensionality reduction techniques, similar to those used in Task 1 of the qCLEF lab, can improve efficiency to some extent. But, it is important to note that in the case of our clustering formulation, the size of the problem mainly depends on the number of documents and not in the number of their features. However, further research in this direction could help overcome hardware limitations.

Finally, fine-tuning the quantum annealing parameters, such as the annealing time and the number of reads, could yield additional improvements in clustering quality. As quantum computing technology continues to evolve, the development of more advanced quantum algorithms and hardware advancements will likely enhance the scalability and efficiency of these methods. In conclusion, our study demonstrates the potential of quantum annealing to enhance clustering performance for IR and RS. The hierarchical model we propose provides a viable solution to the challenges posed by current quantum hardware. By continuing to refine hierarchical methods, investigating advanced preprocessing techniques, and leveraging advances in quantum computing technology, we can harness the power of quantum annealing to tackle the complex challenges of clustering large-scale data in the field of IR and RS.

References

[1] A. K. Jain, Data clustering: 50 years beyond k-means, Pattern Recognition Letters 31 (2010) 651–666. URL: https://www.sciencedirect.com/science/article/pii/S0167865509002323. doi:https://doi.org/10.1016/j.patrec.2009.09.011, award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[2] C. Carpineto, S. Osiński, G. Romano, D. Weiss, A survey of web clustering engines, ACM Comput. Surv. 41 (2009). URL: https://doi.org/10.1145/1541880.1541884. doi:10.1145/1541880.1541884.

[3] K. Djouzi, K. Beghdad-Bey, A review of clustering algorithms for big data, in: 2019 International Conference on Networking and Advanced Systems (ICNAS), 2019, pp. 1–6. doi:10.1109/ICNAS.2019.8807822.

[4] S. Pitafi, T. Anwar, Z. Sharif, A taxonomy of machine learning clustering algorithms, challenges, and future realms, Applied Sciences 13 (2023). URL: https://www.mdpi.com/2076-3417/13/6/3529. doi:10.3390/app13063529.

[5] T. Kadowaki, H. Nishimori, Quantum annealing in the transverse ising model, Phys. Rev. E 58 (1998) 5355–5363. URL: https://link.aps.org/doi/10.1103/PhysRevE.58.5355. doi:10.1103/PhysRevE.58.5355.

[6] T. Albash, D. A. Lidar, Adiabatic quantum computation, Rev. Mod. Phys. 90 (2018) 015002. URL: https://link.aps.org/doi/10.1103/RevModPhys.90.015002. doi:10.1103/RevModPhys.90.015002.

[7] K. Kurihara, S. Tanaka, S. Miyashita, Quantum annealing for clustering, in: Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence, 2009.

[8] V. Kumar, G. Bass, C. Tomlin, J. Dulny, Quantum annealing for combinatorial clustering, Quantum Information Processing 17 (2018) 39. URL: https://doi.org/10.1007/s11128-017-1809-2. doi:10.1007/s11128-017-1809-2.

[9] C. Bauckhage, N. Piatkowski, R. Sifa, D. Hecker, S. Wrobel, A qubo formulation of the k-medoids problem, 2019. URL: https://publica.fraunhofer.de/handle/publica/405469.

[10] D. Arthur, P. Date, Balanced k-means clustering on an adiabatic quantum computer, Quantum Information Processing 20 (2021) 294. URL: https://doi.org/10.1007/s11128-021-03240-8. doi:10.1007/s11128-021-03240-8.

[11] H. Asaoka, K. Kudo, Nonnegative/binary matrix factorization for image classification using quantum annealing, Scientific Reports 13 (2023) 16527. URL: https://doi.org/10.1038/s41598-023-43729-z. doi:10.1038/s41598-023-43729-z.

[12] P. Bermejo, R. Orús, Variational quantum and quantum-inspired clustering, Scientific Reports 13 (2023) 13284. URL: https://doi.org/10.1038/s41598-023-39771-6. doi:10.1038/s41598-023-39771-6.

[13] A. Pasin, M. Ferrari 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, Springer Nature Switzerland, Cham, 2023, pp. 97–108.

[14] 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.

[15] 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.

[16] P. Hauke, H. G. Katzgraber, W. Lechner, H. Nishimori, W. D. Oliver, Perspectives of quantum annealing: methods and implementations, Reports on Progress in Physics 83 (2020) 054401. URL: https://dx.doi.org/10.1088/1361-6633/ab85b8. doi:10.1088/1361-6633/ab85b8.

[17] S. Yarkoni, E. Raponi, T. Bäck, S. Schmitt, Quantum annealing for industry applications: introduction and review, Reports on Progress in Physics 85 (2022) 104001. URL: https://dx.doi.org/10.1088/1361-6633/ac8c54. doi:10.1088/1361-6633/ac8c54.

[18] M. W. Johnson, M. H. S. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Johansson, P. Bunyk, E. M. Chapple, C. Enderud, J. P. Hilton, K. Karimi, E. Ladizinsky, N. Ladizinsky, T. Oh, I. Perminov, C. Rich, M. C. Thom, E. Tolkacheva, C. J. S. Truncik, S. Uchaikin, J. Wang, B. Wilson, G. Rose, Quantum annealing with manufactured spins, Nature 473 (2011) 194–198. URL: https://doi.org/10.1038/nature10012. doi:10.1038/nature10012.

[19] T. Stollenwerk, B. Oauthor=Andrea Pasin and Maurizio Ferrari Dacrema and Paolo Cremonesi and Nicola Ferro, 'Gorman, D. Venturelli, S. Mandrà, O. Rodionova, H. Ng, B. Sridhar, E. G. Rieffel, R. Biswas, Quantum annealing applied to de-conflicting optimal trajectories for air traffic management, IEEE Transactions on Intelligent Transportation Systems 21 (2020) 285–297. doi:10.1109/TITS.2019.2891235.

[20] F. Neukart, G. Compostella, C. Seidel, D. von Dollen, S. Yarkoni, B. Parney, Traffic flow optimization using a quantum annealer, Frontiers in ICT 4 (2017). URL: https://www.frontiersin.org/articles/10.3389/fict.2017.00029. doi:10.3389/fict.2017.00029.

[21] G. Pilato, F. Vella, A survey on quantum computing for recommendation systems, Information 14 (2023). URL: https://www.mdpi.com/2078-2489/14/1/20. doi:10.3390/info14010020.

[22] I. Chakrabarty, S. Khan, V. Singh, Dynamic grover search: applications in recommendation

systems and optimization problems, Quantum Information Processing 16 (2017) 153. URL: https://doi.org/10.1007/s11128-017-1600-4. doi:10.1007/s11128-017-1600-4.

[23] R. Bhagawati, T. Subramanian, An approach of a quantum-inspired document ranking algorithm by using feature selection methodology, International Journal of Information Technology 15 (2023) 4041–4053. URL: https://doi.org/10.1007/s41870-023-01543-w. doi:10.1007/s41870-023-01543-w.

[24] M. Ferrari Dacrema, F. Moroni, R. Nembrini, N. Ferro, G. Faggioli, P. Cremonesi, Towards feature selection for ranking and classification exploiting quantum annealers, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 2814–2824. URL: https://doi.org/10.1145/3477495.3531755. doi:10.1145/3477495.3531755.

[25] R. Nembrini, M. Ferrari Dacrema, P. Cremonesi, Feature selection for recommender systems with quantum computing, Entropy 23 (2021). URL: https://www.mdpi.com/1099-4300/23/8/970. doi:10.3390/e23080970.

[26] M. Ferrari Dacrema, N. Felicioni, P. Cremonesi, Optimizing the selection of recommendation carousels with quantum computing, in: Proceedings of the 15th ACM Conference on Recommender Systems, RecSys '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 691–696. URL: https://doi.org/10.1145/3460231.3478853. doi:10.1145/3460231.3478853.

[27] B. Fauseweh, Quantum many-body simulations on digital quantum computers: State-of-the-art and future challenges, Nature Communications 15 (2024) 2123. URL: https://doi.org/10.1038/s41467-024-46402-9. doi:10.1038/s41467-024-46402-9.