VEnOM: A Vector Embedding Operator Modelling Framework

Andreas Loizou
School of Electrical and Computer Engineering/ Database and Knowledge Systems Lab
National Technical University of Athens
Athens, Greece
antreasloizou@mail.ntua.gr

Nikolaos Andriotis
School of Electrical and Computer Engineering/ Database and Knowledge Systems Lab
National Technical University of Athens
Athens, Greece
nandriotis@mail.ntua.gr

Dimitrios Tsoumakos
School of Electrical and Computer Engineering/ Database and Knowledge Systems Lab
National Technical University of Athens
Athens, Greece
dtsouma@mail.ntua.gr

Abstract

The massive increase in available data sources among organisations has generated new challenges for scalable and efficient data analytics. Selecting the highest-quality datasets for a specific analytics task can be cumbersome, especially when the number of available inputs is very large. In this demonstration, we present *VEnOM*, a modular modelling system that addresses this challenge: Through *VEnOM*, users have high-precision predictions of the result of an analytics operator for a random input dataset at hand, without actually executing it. *VEnOM* leverages dataset similarity and adaptive modelling in order to accurately infer operator outputs for heterogeneous such operators and dataset types through its modular design. In this demonstration, we showcase the modelling of multiple operators from the domain of machine learning and graph analytics that receive tabular and graph datasets as input.

ACM Reference Format:
Andreas Loizou, Nikolaos Andriotis, and Dimitrios Tsoumakos. 2025. VEnOM: A Vector Embedding Operator Modelling Framework. In *The 19th ACM International Conference on Distributed and Event-based Systems (DEBS '25), June 10–13, 2025, Gothenburg, Sweden.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3701717.3733227

1 INTRODUCTION

Big data applications face a rapid evolution in the volume, variety, and velocity of processed data [6]. The selection of "right" data, that is, by examining the content instead of the size of your data, improves the quality and performance of big data analytics. Data-centric Artificial Intelligence (AI) [16] has drawn increasing interest from researchers. This approach focuses on the quality, context, and structure of data. By doing so, it enhances data quality and improves the performance of analytical or machine learning (ML) algorithms. As the data sources and datasets available in the data centers of the organizations expand, a challenge emerges: We must determine which datasets should be selected and passed into analytic workflows to produce the most accurate results and predictions. Manually selecting high-quality datasets according to the organisation's requirements, which involves having human experts process all available datasets, results in low prediction performance and is time-consuming. Data-centric AI techniques [8], have been proposed to assist organisations to automate the selection of high-quality datasets, and improve overall performance.

In previous work [2], a methodology and respective system that assisted in analytics operator modelling was presented. This work shifted the complexity of near-exhaustive executions towards the creation of a similarity matrix among all the datasets. Through multidimensional scaling (MDS), datasets were mapped into a low-dimensional space; a small subset is then used as a training set for the system to model different possible operators.

While this work showed relatively low prediction errors and significant speed-ups, the method suffered from two serious shortcomings: Firstly, the similarity matrix construction requires a similarity function that is effective for a *single* data metric (i.e., data distribution, order, size, etc.). Secondly, for the consequent MDS to be efficient, datasets are projected into a low-dimensional dimensional space (usually less than 7-d), which further decreases the expressiveness and flexibility of the models. In contrast, *VEnOM* introduces dataset vector representations produced from embeddings according to the data types involved [10]. Embeddings convert dataset objects into complex representations that capture inherent properties and relationships between the data. As a result, the proposed system is able to unify and optimize the analytics modelling processing by incorporating variable and accurate lower-dimensional data representations produced by state-of-the-art embeddings specific to the type of datasets at hand. Indeed, via utilizing existing embedding methods for documents, tables, images, graphs, etc., *VEnOM* can execute its modelling pipeline in a unified and performant manner.

To demonstrate *VEnOM*'s aforementioned ability, we incorporate two diverse data types: Numerical tabular datasets and graphs. The significance of graph analytics has been well-established in numerous application domains, such as social network analysis, bioinformatics, cybersecurity, and recommendation systems. With *VEnOM*'s modular design and framework, users are able to approximate a possibly heavy computational analysis by speeding up the exact evaluation over all graphs.

Modern ML and AI techniques largely aim at improving data analysis using two major tools: *Data Vectorisation* and *Modelling*. Inspired by both, this work proposes *VEnOM*, an operator modelling system that is able to effectively predict analytics operator outputs via their vector embedding representations. *VEnOM* features a modular design, allowing different technologies and algorithms to be incorporated into its pipeline. Given a (possibly large) number of

DEBS '25, Gothenburg, Sweden
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1332-3/25/06
https://doi.org/10.1145/3701717.3733227

datasets, *VEnOM* first vectorizes them with the appropriate embedding method. Resulting vectors can be stored locally or inside a specialized Vector Database (Qdrant [1]) that enhances system efficiency and scalability. For any given dataset for which we wish to know an operator's outcome, a similarity search in the available vectors is performed. The output of this search is used as a training set in order to model the operator's output with high-quality, relevant to the most similar input data available. Dataset embeddings can be used for modelling different operators, as the process is purely data- and not algorithm-dependent: As such, this step of our system is entirely *operator-agnostic*. Moreover, the embedding size, similarity algorithm, the size of the resulting similarity search and, of course, the analytics operator to be modelled are all user-defined and can be modularly realized. We can thus briefly summarize the contributions of this system as follows:

- We describe our modelling framework which aims at providing accurate predictions of analytical task outcomes, given a plethora of available datasets to be analyzed. Our method utilizes vector embedding representations in order to satisfy meaningful, efficient and largely operator- and data-type-agnostic modelling.
- We showcase a prototype implementation of our system [1] which is highly modular, being able to integrate with different storage, embedding, similarity and modelling back-ends. In this demonstration, we show how *VEnOM* is able to seamlessly operate over tabular numerical and graph data and tasks, integrating with a Vector Database and offering a plethora of customization knobs for its users.

2 RELATED WORK

Our work relates to the areas of operator modeling and dataset vectorization. Prior works [2, 4] aimed to model analytics operators by creating a similarity matrix among datasets (both tabular and graph-based) and projecting it into a lower-dimensional space. Subsequently, an operator was modeled using a small random subset of these datasets. In contrast, our proposed approach, *VEnOM*, specifically focuses on dataset vectorization and employs similarity search techniques to select datasets of the highest quality for modeling an operator. The goal of dataset vectorization is to transform a dataset into a vector of fixed size suitable for input into a machine learning model. Similar to our approach, Table2vec [17] projects tabular datasets into vector embeddings by considering not only intrinsic data characteristics but also other table elements such as metadata, captions, column headers, and entity embeddings. However, *VEnOM* utilizes a custom deep learning architecture explicitly tailored to vectorize predominantly numerical tabular datasets.

In the case of graph data, significant progress has been made in developing effective embedding techniques. Many state-of-the-art methods have drawn inspiration from the Skip-Gram model (word2vec) [11], which learns word representations based on their co-occurrence within sentence windows. This idea has been extended to graphs, leading to methods such as DeepWalk [14], LINE [15], and node2vec [5], which generate node-level embeddings by leveraging random walks and network structures. Building on the success of word2vec, the graph2vec model [13] was introduced to learn data-driven, distributed representations of entire graphs. Unlike node-level methods, graph2vec generates a single embedding vector per graph, enabling its use in a variety of downstream tasks such as graph classification, clustering, and as input for supervised representation learning. In this work, we integrate the graph2vec methodology into the VEnOM framework to enable scalable and expressive graph-level learning.

3 SYSTEM OVERVIEW

In the following section we provide a brief overview of *VEnOM's* architecture and our framework's overall pipeline.

3.1 Processing Framework

VEnOM selects the most relevant datasets from a data repository (referred to as a "data lake") to model real-world operators and enhance performance. Initially, each dataset D_z in the data lake D undergoes embedding-based vectorization. Using a similarity search, *VEnOM* takes a target dataset D_o (also vectorized) as input, for which we wish to predict an operator's outcome, and identifies the optimal subset of datasets D_r for modeling the operator Φ. This subset D_r is then employed to build a predictive model for Φ (e.g., using SVM, ANN, Decision Trees, etc.). The resulting model can be subsequently queried to predict the outcome $\Phi(D_o)$. Figure 1 illustrates the processing stages of our framework, which comprises three distinct phases as described in further detail below:

3.1.1 Dataset Vectorisation. *VEnOM* is designed to handle a variety of datasets, including structured (tabular) and semi-structured (graph) types. Utilizing a vector embedding approach, the framework extracts key features from each dataset and represents them as vectors D_z in a lower-dimensional space of dimension k. Users can select from various vectorization models and specify the dimension of the embedding vectors. Different models can be applied based on dataset types, allowing analysts to balance between execution time and operator accuracy according to vector dimensionality. For tabular numerical datasets, *VEnOM* employs a custom model that transforms datasets into vector embeddings of dimension k [10]. *VEnOM* also provides functionality to store these vectors locally or in an open-source Vector Database (the Qdrant Vector DB [1]). Additionally, our system can handle graph datasets by leveraging the state-of-the-art *graph2vec* embeddings [13]. Graph2vec constructs a vocabulary from rooted subgraphs or neighborhoods around each node and learns the representations of entire graphs through the doc2vec skip-gram training process.

3.1.2 Similarity estimation. Once vector representations are stored, *VEnOM* can perform various similarity search algorithms to identify datasets within the data lake D that are most similar to the target dataset D_o. This phase involves selecting one of several similarity methods to determine the most suitable datasets for accurate operator modeling. *VEnOM* is seamlessly integrated with the Qdrant Vector DB [1]. As such, the similarity search methods supported by Qdrant can also be utilized within the *VEnOM* framework. Integrating a dedicated Vector DB enhances both the accuracy of similarity searches and the efficiency of the process, as such systems are optimized for scalability and performance.

[1] https://github.com/aloizo03/VEnOM-A-Vector-Embedding-Operator-Modelling-Framework-

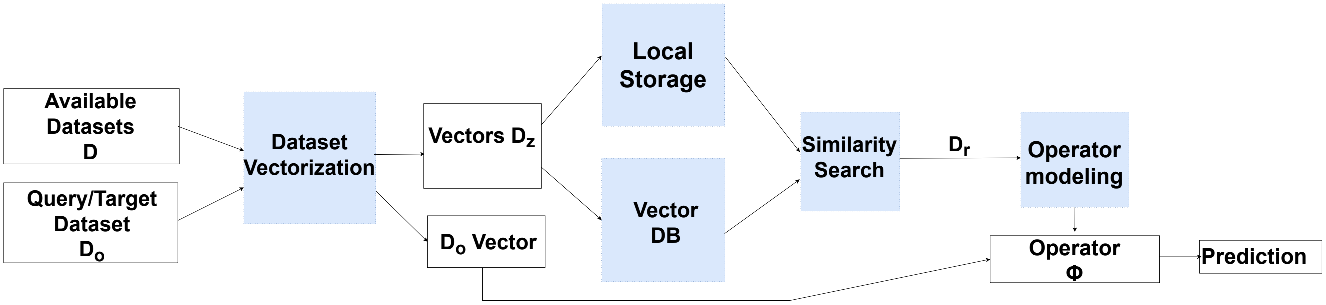


Figure 1: *VEnOM* framework architecture

3.1.3 Operator Modelling. Operator modelling is conducted by selecting one of the available algorithms to create and export an operator model to a specified filesystem directory. *VEnOM* offers a broad range of operator modelling algorithms, including machine learning methods (e.g., linear/logistic regression, MLP, etc.), clustering algorithms (e.g., DBSCAN, K-Means), and time series models (e.g., ARIMA, Holt-Winters). For graph datasets, *VEnOM* currently supports modelling the following graph-based operators: Betweenness Centrality (BC), Edge Betweenness Centrality (EBC), Closeness Centrality (CC), Eigenvector Centrality (EC), and PageRank (PR). To model an operator, the user must select the appropriate algorithm and specify the target directory where the model will be stored for future reuse.

3.2 Initial Results

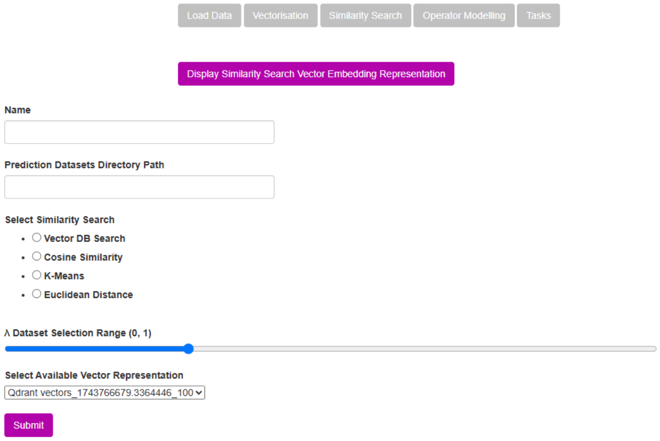
Our initial evaluation of the framework assesses both prediction accuracy and runtime efficiency. Prediction accuracy is measured using the Mean Absolute Error (MAE) loss function. Runtime efficiency is assessed through Amortized Speedup, which incorporates the one-time cost of vectorizing the entire data lake and accounts for reuse across multiple operator modeling tasks, as discussed in [10]. For the evaluation presented here, we focus on two datasets: the Household Power Consumption (HPC) dataset [7], which comprises 401 datasets with measurements of electric power consumption in a household in Sceaux, France; and the Weather dataset [3], which contains hourly weather measurements from 36 U.S. cities spanning the years 2012 to 2017. For the HPC dataset, we modelled the Linear Regression (LR) operator using three vector dimension sizes (100, 200, and 300), employing Euclidean distance as the similarity metric. For the Weather dataset, we modelled the Multilayer Perceptron (MLP) operator across the same set of vector dimensions (100, 200, and 300), and employed cosine similarity for similarity search.

Table 1: *VEnOM* evaluation results

Dataset Name	Operator	Similarity Metric	Vector Size	MAE	Amortized Speedup
Household Power Consumption	Linear Regression	Euclidean	100	8.02	3.85
			200	5.29	2.45
			300	5.24	1.99
Weather	Multilayer Perceptron	Cosine	100	8.28	14.64
			200	12.76	4.5
			300	4.3	1.64
COLLAB	PageRank	K-Means	128	0.10	16.98

For graph data, we conducted experiments using the PageRank operator over the COLLAB dataset provided by [12]. The COLLAB network is a scientific collaboration dataset derived from three public collaboration sources [9], specifically High Energy Physics, Condensed Matter Physics, and Astro Physics. The dataset comprises 5,000 graphs, with an average of 74.49 nodes and 2,457.78 edges per graph. Table 1 presents our findings in the three cases described.

4 DEMONSTRATION SCENARIO



Name

Prediction Datasets Directory Path

Select Similarity Search

- ○ Vector DB Search
- ○ Cosine Similarity
- ○ K-Means
- ○ Euclidean Distance

λ Dataset Selection Range (0, 1)

Select Available Vector Representation

Qdrant vectors_1743766679.3364446_100

Figure 2: Windows for similarity search

In this section we follow the steps taken in our system's UI in order to demonstrate *VEnOM*. The user may upload various collections of datasets (tabular data or graphs) to the system using the web UI, and each dataset in the collection is vectorized into a dimension selected by the user. Subsequently, the vectors are saved and reused in later pipeline phases. Our framework is deployed in a AWS EC2 virtual machine with eight 2nd Generation Intel Xeon Scalable Processors (Cascade Lake P-8259CL), 32GB RAM memory, and one T4 Tensor Core GPU 16GB.

When using our system, an analyst can follow the steps as described in Section 3.1. In Figure 2, the individual phases of the system are shown with grey colour. The *"Load Data"* tab enables the user to upload a dataset repository. In the *"Vectorization"* tab, the user is prompted to select a vectorization method as well as the dimension of the embedding. Then, in the *"Similarity Search"* tab, as shown in Figure 2 for demonstration purposes, the most relevant to the input (query) datasets are identified. In the final step, the *"Operator Modelling"* tab will model and output an operator that the user has selected. Additionally, the *"Task"* tab allows users to view all asynchronous process phases and monitor the progress of each assigned task. When a new pipeline task is issued, the back-end processes the request asynchronously as the user navigates the application's UI. All executing pipeline tasks and their status can be tracked from the system tabs as mentioned above.

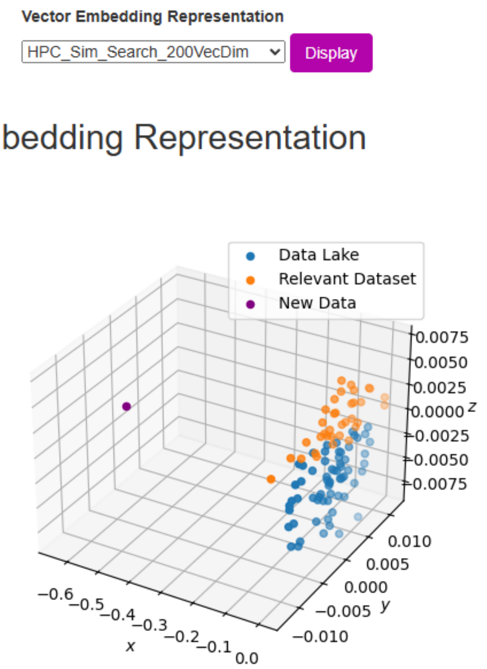


Figure 3: 3-D Similarity Search Dataset Vector Embedding Representation

As the similarity estimation plays a crucial role in VEnOM's framework, we illustrate the output of that tab in Figure 3. After the analyst chooses all the necessary options in the *"Similarity Search"* tab, the vector embedding representation of each dataset is projected from the k-dimensional plane to a 3-d plane. Looking at Figure 3, with orange colour the user can overview all the available datasets from the data lake. With purple colour, the analyst distinguishes her input dataset and with blue, the datasets output from the similarity search operation. The closer a dataset is, the more intense its colouring.

Having uploaded a dataset, calculated the similarity search, and identified by means of a similarity search the highest quality datasets for modelling, the analyst finally chooses an operator/algorithm to infer its output. The prediction is promptly shown, including the error of the modelled operator, graphs with various statistical error/accuracy measurements such as MaPE, RMSE, MAE, and the speedup of our technique. As a result, the user can do a more in-depth examination of the modelled operator and the chosen parameters.

Acknowledgments

This work is partially supported by the project RELAX-DN, funded by the European Union under Horizon Europe 2021-2027 Framework Programme Grant Agreement number 101072456.

References

[1] 2021. *Qdrant - Vector Database — qdrant.tech.* [Accessed 07-12-2024].

[2] Tasos Bakogiannis, Ioannis Giannakopoulos, Dimitrios Tsoumakos, and Nectarios Koziris. 2019. Apollo: A dataset profiling and operator modeling system. In *Proceedings of the 2019 International Conference on Management of Data.* 1869–1872.

[3] Beniaguev David. 2020. Weather Dataset.

[4] Ioannis Giannakopoulos, Dimitrios Tsoumakos, and Nectarios Koziris. 2018. A Content-Based Approach for Modeling Analytics Operators. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management.* 227–236.

[5] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16).* Association for Computing Machinery, New York, NY, USA, 855–864. https://doi.org/10.1145/2939672.2939754

[6] Benjamin T Hazen, Christopher A Boone, Jeremy D Ezell, and L Allison Jones-Farmer. 2014. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics* 154 (2014), 72–80.

[7] Georges Hebrail and Alice Berard. 2006. Individual Household Electric Power Consumption. UCI Machine Learning Repository.

[8] Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. 2024. Data-centric artificial intelligence. *Business & Information Systems Engineering* (2024), 1–9.

[9] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (Chicago, Illinois, USA) *(KDD '05).* Association for Computing Machinery, New York, NY, USA, 177–187. https://doi.org/10.1145/1081870.1081893

[10] Andreas Loizou and Dimitrios Tsoumakos. 2025. Data Analysis Prediction over Multiple Unseen Datasets: A Vector Embedding Approach. *arXiv preprint arXiv:2502.17060* (2025).

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13).* Curran Associates Inc., Red Hook, NY, USA, 3111–3119.

[12] Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. 2020. TUDataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020).* arXiv:2007.08663 www.graphlearning.io

[13] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. 2017. graph2vec: Learning Distributed Representations of Graphs. arXiv:1707.05005 [cs.AI] https://arxiv.org/abs/1707.05005

[14] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, New York, USA) *(KDD '14).* Association for Computing Machinery, New York, NY, USA, 701–710. https://doi.org/10.1145/2623330.2623732

[15] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) *(WWW '15).* International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1067–1077. https://doi.org/10.1145/2736277.2741093

[16] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158* (2023).

[17] Li Zhang, Shuo Zhang, and Krisztian Balog. 2019. Table2vec: Neural word and entity embeddings for table population and retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval.* 1029–1032.