Apollo: A Dataset Profiling and Operator Modeling System

Tasos Bakogiannis[1], Ioannis Giannakopoulos[1], Dimitrios Tsoumakos[2] and Nectarios Koziris[1]
[1]Computing Systems Laboratory, School of ECE, National Technical University of Athens, Greece
{abk,ggian,nkoziris}@cslab.ece.ntua.gr
[2]Department of Informatics, Ionian University, Corfu, Greece
dtsouma@ionio.gr

ABSTRACT

The rapidly increasing amount of available data has created invaluable business opportunities but also new challenges. The focus on content-driven analytics is shifting attention from optimizing operators and systems to handle massive data sizes, to intelligent selection of those datasets that maximize the business competitive advantage. To date, there exists no efficient method to quantify the impact of numerous available datasets over different analytics tasks – a thorough execution over every input would be prohibitively expensive. In this demonstration, we present *Apollo*, a data profiling and operator modeling system that tackles this challenge. Our system quantifies dataset similarities and projects them into a low-dimensional space. Operator outputs are then estimated over the entire dataset, utilizing similarity information with Machine Learning and a small sample of actual executions. During the demo, attendees will be able to model and visualize multiple analytics operators over datasets from the domains of machine learning and graph analytics.

ACM Reference Format:
Tasos Bakogiannis, Ioannis Giannakopoulos, Dimitrios Tsoumakos and Nectarios Koziris. 2019. Apollo: A Dataset Profiling and Operator Modeling System. In *2019 International Conference on Management of Data (SIGMOD '19), June 30-July 5, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3299869.3320220

SIGMOD '19, June 30-July 5, 2019, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5643-5/19/06…$15.00
https://doi.org/10.1145/3299869.3320220

1 INTRODUCTION

As Big Data technologies mature and evolve, efforts are placed on areas not solely related to data size. An increasing number of research works make the case for the "Right Data" [1], where content rather than size is the critical factor for data analysis workflows. The plethora of available data sources for content-sensitive analytics leads to the need to identify *high impact* data, i.e., intelligence that has the best potential of driving strategic decisions. Thus, data scientists have to decide on which datasets should be applied to each given analytics workflow independently. Given the increasing complexity of modern workflows and the plethora and diversity of available data operators, evaluating the utility of immense numbers of inputs for a given workflow is prohibitively expensive.

As a motivating example, consider the case of derivative pricing theory [6]. Analysts need to consider a multitude of Credit Default Swaps (CDS) time series for different economic entities. These are provided as input to mathematically complex operators so that financial indicators, i.e., Value Adjustments (xVA), are extracted that quantify the credit, funding and financial costs an institution faces during derivative transactions. Selecting the appropriate CDS datasets for extracting the respective xVAs for an entity is of key importance for the indicator's accuracy. Thus, an analyst needs to exhaustively compute the aforementioned operators for all CDS datasets in order to select those that present certain characteristics that make them more suitable for specific entities and maximize accuracy.

In the domain of graph analytics, consider a dataset consisting of many citation graphs. We wish to identify the graphs with the most well-connected citations that contain highly-cited papers. The clustering coefficient, a good measure of neighborhood connectivity, would have to be computed for all the graphs in order to allow the identification of the top-k such graphs. To quantify the importance of each paper, we consider a centrality measure such as betweenness centrality. Consequently, we would have to compute

the maximum betweenness centrality score for each citation graph and combine the results with those obtained from the analysis based on the clustering coefficient. Yet, this is a daunting task due to the operators' complexity and the number of executions required.

To facilitate dataset analysis, two complementary directions have been suggested: *Data Integration* and *Data Exploration*. Drawing inspiration from both, this work proposes *Apollo*[1], an operator-agnostic dataset profiling tool. In order to avoid the exhaustive execution of the available operators over each dataset, our work assesses the similarity between datasets employing similarity measures that correlate with the behavior of each operator. Based on the dataset relationships, we infer knowledge about them. Each dataset is then projected to a point in a low-dimensional metric space that preserves the dataset similarities in the form of distances between the respective points. Using this metric space, datasets are sampled and the operator to be modeled is applied to each of the samples. Finally, *Apollo* models the given operator's output for all datasets, using Neural Networks, based on the output samples and the all-pairs dataset similarities already computed.

Interestingly, the construction of the dataset metric space is *operator-agnostic*. As a result, *Apollo* can re-use it for modeling different operators. Hence, our work shifts the effort towards measuring the dataset inter-relationships which we leverage to model different operators instead of computing their outputs for each dataset. The contribution of this work is threefold:

- We present *Apollo*, an operator-agnostic dataset profiling framework [5] that aims at modeling operator output for a collection of datasets. To do so, it computes the similarity among datasets, constructs a dataset space that reflects their properties and models the output of the applied operators utilizing machine learning.
- We offer an open source implementation in Go [4], distributable as a Docker image. The implementation includes the dataset profiling pipeline, visualizations of the dataset similarities, the dataset metric space, etc, as well as the extraction of the generated ML models for integration with other systems.
- We evaluate the accuracy and efficiency of the proposed system using operators and datasets from different domains. Specifically, popular ML operators (e.g., clustering, linear regression, time-series prediction, etc) and graph measures (e.g., various centralities, PageRank, diameter, etc) are evaluated over both real and synthetic datasets.

[1]According to Greek mythology, *Apollo* was one of the Olympian deities who was worshiped at the famous oracle at the city of Delphi.

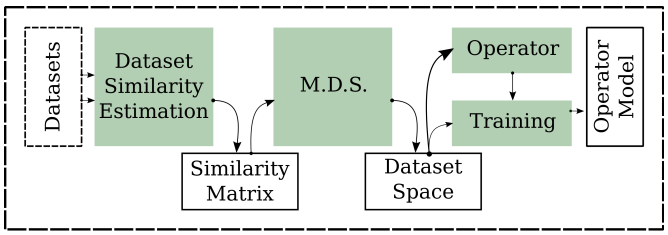


Figure 1: *Apollo*'s Processing Pipeline

2 SYSTEM OVERVIEW

In this Section we give a brief overview of *Apollo*'s architecture and processing pipeline.

2.1 Processing Pipeline

Apollo is based on the observation that datasets with similar statistical properties affect a wealth of real world operators in similar ways and, hence, lead them to produce similar outputs. To this end, Apollo is designed to assess the similarity between different datasets in the light of different statistical properties (including the *distribution*, the *size* and the *order* in tabular data), project this knowledge to a low-dimensional, easy-to-inspect space and, finally, model operator output for all datasets using Neural Networks. Figure 1 provides the overview of the processing pipeline. Let us now discuss each step in more detail:

2.1.1 Similarity Estimation: Apollo can accept and operate on multiple collections of datasets available on storage. For each collection, the user first computes the similarity matrix of the datasets in the collection. By choosing an available similarity measure, uploading a custom one or even combining two different measures, the analyst can create a matrix of all-pairs similarities between the datasets of the collection and store it for further use.

The ability to specify the similarity measure enables the analyst to re-interpret a dataset based on his needs. For example, calculating the similarity between graphs is a difficult task, comparing distributions, on the other hand, can be done much more efficiently. To this end, *Apollo* allows us to introduce a similarity measure that assesses the similarity between graphs by calculating their degree distributions and comparing the distributions instead.

2.1.2 Dataset Space Construction: Given the possible high dimensionality of the dataset space, *Apollo* can perform Multidimensional Scaling (MDS) having as input the already generated similarity matrix interpreted as a distance matrix. The purpose of this step is twofold: It enables the analyst to gain insight and intuition on the dataset space by reducing the dimensions to 2 or 3 and then visualizing the result. In addition, it significantly contributes to the efficiency of the modeling step since most Machine Learning algorithms require the coordinates of the input data points rather than their inter-similarities.

2.1.3 Operators and Modeling: For *Apollo*, an operator is any executable that can be run against a dataset and produce a numerical value as a result. This definition can encompass a broad collection of operators including ML (DBSCAN, linear regression, etc), statistical (avg, sum, etc) or specific to the dataset type, e.g, for graph data, centralities (betweenness, closeness, etc), spectrum related (PageRank, spectral radius, etc) or general graph properties (e.g., diameter). To model such an operator, the user has to import the executable to the system and run it for a random sample of datasets. He then chooses one of the available Neural Network configurations and trains a model based on the sampled operator outputs and the dataset space created.

2.2 Experimental Evaluation

For a thorough evaluation of our system in terms of accuracy and efficiency, i.e., speedup compared to a brute force alternative, we refer the reader to the works in [2, 5], where an extensive evaluation for a variety of dataset types is performed. For the purpose of this demonstration, we focus on two sets of datasets, a set of 973 ego graphs from Twitter (*TW*) [8] and a set of 1442 datasets with daily household power consumption measurements (*HPO*) from a household in Denmark [9]. For TW, we consider Betweenness Centrality (bc) and PageRank (pr), two widely used node centrality measures generalized to the graph level through Freeman's method [3]. For HPO, we model the Average (avg) of each dataset and the number of clusters created after performing a DBSCAN (dbs). We present *MdAPE*, as a measure of accuracy, and speedup results for two different sampling ratios (p=5%,10%) in Table 1.

Table 1: Modeling Errors and Speedups

Dataset	Operator	MdAPE (%)		Speedup ×	
		p=5%	p=10%	p=5%	p=10%
TW	bc	17.8	17.5	13.0	7.8
	pr	9.2	7.7	13.2	7.9
HPO	avg	1.3	1.2	3.93	3.4
	dbs	14.6	14.1	8.3	6.23

2.3 Visualizations, Evaluation and Model Export

Acknowledging data visualization as a very effective way to develop intuition about datasets, *Apollo* provides the analyst with a number of visualizations. The similarity matrix can be visualized as a heat map, the dataset space can be reduced to 2 or 3 dimensions and be displayed in an interactive chart. In the same chart, the user can visualize the results of the modeling procedure, i.e., the samples and approximations for each operator. Additionally, in order for the analyst to be able to evaluate the accuracy of an operator's model, a number of measures are provided and calculated each time a new model is trained. Finally, following the model training and evaluation, the user can export an operator's model to incorporate it into an analytics pipeline.

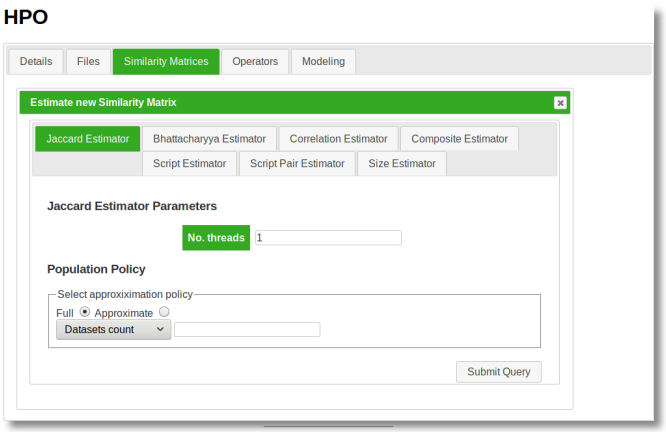


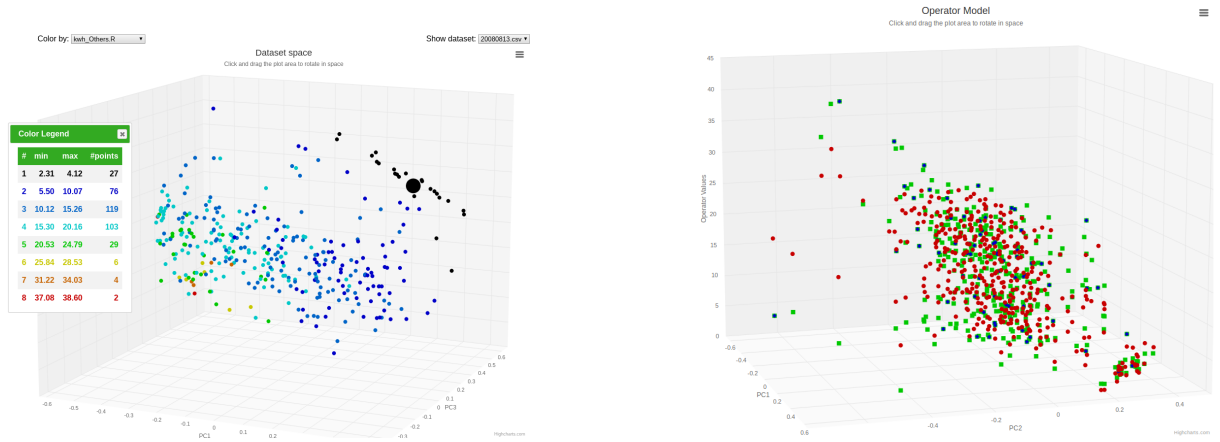
Figure 2: Dialog for similarity estimation

3 DEMONSTRATION SCENARIO

We demonstrate *Apollo* using the HPO dataset presented in Section 2.2. The user can upload a collection of datasets to the system through the web UI and choose from a list of similarity measures, as in Figure 2, in order to create similarity matrices for the collection of datasets.

The steps an analyst takes when using the system, described in Section 2.1, can be broadly represented by the system's tabs (Figure 2): *Files* for dataset uploads, *Similarity Matrices* for similarity matrix and MDS operations, *Operators* for loading new operators and *Modeling* where the models are trained for specific operators and the results are

Figure 3: Similarity Matrix



(a) 3-d Dataset Space Projection **(b) 3-d Projection of Modeling Operator Outputs**

Figure 4: 3-d Projections

visualized. When launching a new task, the backend asynchronously processes the request while the user can browse the application. All running tasks and their status can be monitored from one of the system's tabs.

Selecting one of the generated similarity matrices, the analyst may choose to visualize its data in the form of an interactive heatmap (Figure 3) that can be sorted based on the distances from a given dataset. Alternatively, it is possible to perform MDS after specifying the number of target dimensions of the dataset space. If the resulting space is 2- or 3-dimensional, it is possible to visualize it in an interactive chart. A 3-*d* representation of the HPO dataset is displayed in Figure 4a. In this dataset space projection, each point represents a dataset. Having calculated a set of operators for these datasets, the points can be colorized based on the selected operator outputs. In Figure 4a, the datasets are colored based on their power consumption calculated as kilowatts per hour: The higher the power consumption, the warmer the point's color.

Having calculated an operator for a ratio of the available datasets, the analyst can now choose a neural network design from the *Modeling* tab and train a model which approximates the given operator. To visually evaluate the accuracy of the model, it is possible to project the approximated values in the same dataset space generated after the dimensionality reduction. For example, in Figure 4b the samples are the diamond shaped blue points, the actual values being the square green points and the approximations are the round red points. *Apollo* also provides screens with a collection of statistical error/accuracy measures like *MAPE*, *MdAPE*, *RMSE*, etc, for a more in-depth evaluation of a trained model.

4 RELATED WORK

Our work relates to the areas of *Data Integration* and *Data Exploration*. Works in *Data Integration*, as outlined in [7], mostly aim at unifying distinct datasets or providing context to a dataset and answering questions over that unified collection. Our work, on the other hand, focuses on the differences between datasets and aims at modeling them. In *Data Exploration* the goal is to identify the key properties of a dataset. Similarly to *Apollo*, in works such as [10] statistical analysis is used on the available datasets. However, our work aims at modeling more complex statistical properties and does not only focus on the generation of data aggregates.

ACKNOWLEDGMENTS

This work is partially supported by European Union's Horizon 2020 RIA programme under GA No 690588, project SELIS.

REFERENCES

[1] Ricardo A. Baeza-Yates. 2013. Big Data or Right Data?. In *Proceedings of the 7th Alberto Mendelzon International Workshop on Foundations of Data Management*.

[2] Tasos Bakogiannis, Ioannis Giannakopoulos, Dimitrios Tsoumakos, and Nectarios Koziris. 2019. Predicting Graph Operator Output over Multiple Graphs. In *ICWE 2019*.

[3] Linton C. Freeman. 1977. A Set of Measures of Centrality Based on Betweenness. *Sociometry* 40, 1 (1977), 35–41.

[4] Ioannis Giannakopoulos. 2017. Apollo. https://github.com/giagiannis/data-profiler.

[5] Ioannis Giannakopoulos, Dimitrios Tsoumakos, and Nectarios Koziris. 2018. A Content-Based Approach for Modeling Analytics Operators. In *CIKM 2018*. 227–236.

[6] Jon Gregory. 2010. *Counterparty credit risk: the new challenge for global financial markets*. Vol. 470. John Wiley & Sons.

[7] Maurizio Lenzerini. 2002. Data Integration: A Theoretical Perspective. In *SIGACT-SIGMOD-SIGART 2002*. 233–246.

[8] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.

[9] Lichman. 2013. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[10] Abdul Wasay, Xinding Wei, Niv Dayan, and Stratos Idreos. 2017. Data Canopy: Accelerating Exploratory Statistical Analysis. In *SIGMOD 2017*. 557–572.