Predicting Scientific Data Access Patterns Using dCache Logs

Julian Bellavita
University of California, Berkeley
jbellavita@berkeley.edu

I. Introduction

The dCache storage management system is a disk cache at Brookhaven National Lab for large high-energy physics (HEP) datasets produced by the ATLAS experiment [1]. Storage space on dCache is limited relative to that of persistent storage devices. Therefore, a heuristic is needed to determine what data should be kept in the cache and to determine what data should be evicted from the cache. A good cache eviction policy is one that seeks to keep frequently accessed data in the cache. However, since data access frequency can change over time, such an eviction policy requires knowledge of future dataset popularity. The goal of this study is to present methods for predicting how many times a dataset will be accessed in the future. We present a novel deep neural network that can forecast how many times a dataset will be accessed one day in the future. We use k-means clustering to group datasets into popularity groups based on their current and future number of accesses. Lastly, we present a set of algorithms that can use a historical dataset access sequence to predict what datasets will be accessed in the future.

II. Methods

This section outlines the methods used to predict dataset popularity. The deep neural network was trained using a dataset containing information for 9 months' worth of dCache transactions. Each record in the dataset contains information regarding a particular dCache dataset on a particular day. The dataset contains the following features: the number of files in a dataset, the number of times a dataset is accessed on the date corresponding to the record, the size of a dataset in bytes, the number of bytes read from a dataset, the number of times a dataset is accessed one day after the date corresponding to the record (this is the feature forecast by the deep neural network), and the type of files found in the dataset (e.g. experiment results, server log data, etc.). The deep neural network was built using PyTorch; it uses 2 dense layers, the Tanh activation function, and the ADAM optimizer.

K-means clustering was performed using Scikit-learn's KMeans class. The datasets are clustered according to their present and future accesses. The optimal k value was determined using the elbow method (see Figure 1).

Some of the file prediction algorithms were obtained from Patra et al. (2010) [2], while Last N Successors (LSN) and Backup Predictor (BP) are novel. All file prediction algorithms were implemented using C++. LSN accepts an integer N and predicts the successor of a dataset to be the most recent successor that was observed at least N times. BP accepts two other file prediction algorithms, a default predictor and a backup predictor. It will attempt to make a prediction with the default predictor, but if the default predictor refuses to make a prediction, it will make a prediction with the backup predictor. Details and parameters for the other algorithms can be found in [2]. Each file prediction algorithm was tested on an access stream of 3125860 dataset accesses extracted from dCache logs. We recorded the total number of predictions each algorithm made and the total number of correct predictions each algorithm made.

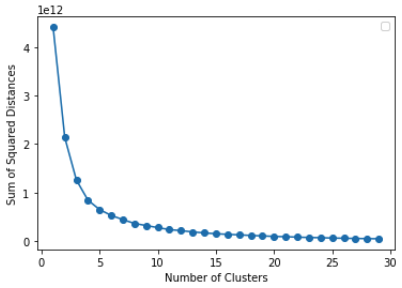

Fig. 1. Sum of squared distances for various k values. Based on this curve, we chose a k value of 5.

III. Results

Figure 2 shows the Mean Absolute Error (MAE) loss on the test dataset. The final loss is 4.6e-8. Figure 3 shows the predicted future access values vs the actual future access values of the training and test data. The data is normalized, hence the small magnitude of the y-axis. Generally, the model performs quite well. Note that although the data in Figure 3 is normalized, we are only interested in the relative values of the predictions in relation to other predictions, not the raw values of the predictions. Therefore, presenting the normalized data is sufficiently informative.

Figure 4 shows the results of clustering the datasets according to their present and future access values. Each point represents a single dataset on a single day. The purple group contains the majority of the datasets, whereas the green and dark blue groups are extremely small. The datasets are distributed in a diagonal manner. K-means clustering reveals that the majority of files have relatively few accesses and that there is a positive correlation between present and future numbers of

accesses. Figure 5 shows the usage history of datasets in each cluster over the course of 4 months. One dataset is chosen from each cluster in Figure 4 as the 'representative' of that cluster. The dataset from each cluster that is accessed for the highest number of days is chosen as the representative of that cluster. Figure 5 demonstrates that there is significant variance in the number of day-to-day accesses.

Figure 6 shows the performance of each file prediction algorithm on a large stream of dCache transactions. BP with Recent Popularity (RP) as the default predictor and LSN as the backup predictor gets the highest number of predictions correct: around 27.2% of the total access stream. RP alone has the highest ratio of correct predictions to total predictions. Around 40% of the predictions RP makes are correct, although it makes fewer predictions than BP.

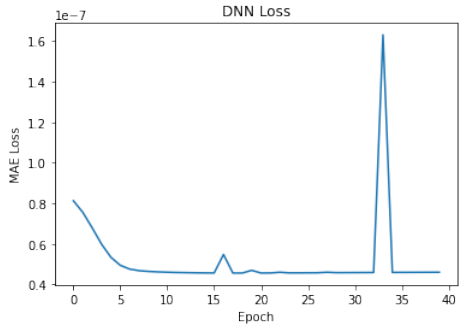

Fig. 2. MAE loss on the test data.

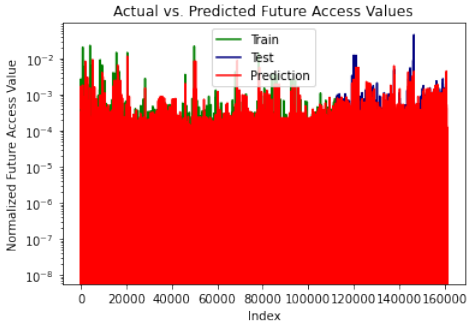

Fig. 3. Predicted vs. actual future access values. Test and train data are shown.

IV. Conclusion

We present various models and algorithms that can predict how many times a dataset will be accessed in the future. Our results indicate that it is possible to forecast dataset popularity in advance. Future work will seek to develop additional file prediction algorithms and to expand the time horizon for which the deep neural network makes predictions.

Acknowledgment

This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and also used resources of the National Energy Research Scientific Computing Center (NERSC). This work was also supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program.

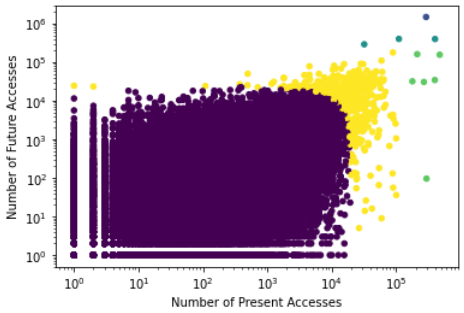

Fig. 4. K-means clustering with k=5.

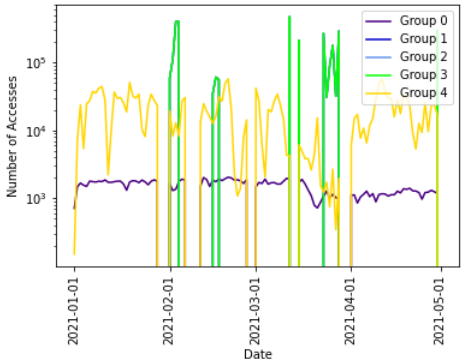

Fig. 5. Dataset popularity over time. Each line is a representative dataset chosen from each cluster in Figure 4.

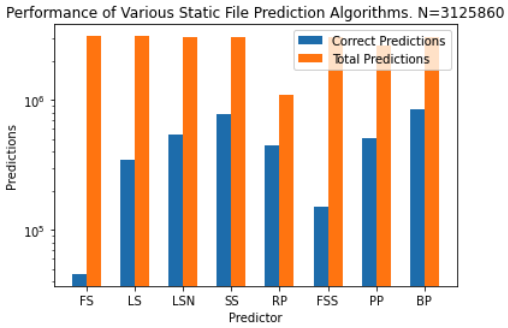

Fig. 6. Performance of each file prediction algorithm on an access stream of 3125860 dataset accesses.

References

[1] Y. Wang, K. Wu, A. Sim, S. Yoo, and S. Misawa, "Access patterns to disk cache for large scientific archive," in *Proceedings of the 2021 on Systems and Network Telemetry and Analytics*, 2020, pp. 37–40.

[2] P. K. Patra, M. Sahu, S. Mohapatra, and R. K. Samantray, "File access prediction using neural networks," *IEEE Transactions on Neural Networks*, vol. 21, no. 6, pp. 869–882, 2010.