Predicting Scientific Dataset Popularity Using dCache Logs

Julian Bellavita[1], Alex Sim (advisor)[2], John Wu (advisor)[2]

[1]University of California, Berkeley, [2]Lawrence Berkeley National Laboratory

ABSTRACT

The dCache installation is a storage management system that acts as a disk cache for high-energy physics (HEP) data. 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, data access frequency can change over time, so this eviction policy requires knowledge of future dataset popularity. We present methods for forecasting the number of times a dataset stored on dCache will be accessed in the future. We present a deep neural network that can predict the number of future dataset accesses with a high degree of accuracy, reporting a final normalized loss of 4.6e-8. We present a set of algorithms that can forecast future dataset accesses given an access sequence. Included in this set are two novel algorithms, Backup Predictor and Last N Successors, that outperform other file prediction algorithms. Findings suggest that it is possible to anticipate dataset popularity in advance, and therefore possible to move popular data into dCache ahead of time.

BACKGROUND INFO

dCache Background

- The ATLAS experiment produces substantial quantities of HEP data.
- dCache has connectivity to a separate high-performance storage system (HPSS), but accessing data stored in the dCache system is faster.
- **Therefore, it is best for frequently accessed data to be stored on dCache.**

MOTIVATION

Forecasting the number of dataset accesses in the future allows movement of popular datasets into the dCache in advance, which will substantially reduce average dataset access latency.

RESEARCH QUESTION

Is it possible to use historical dataset information to predict how many times a dataset will be accessed in the future?

METHODS

Deep Neural Network

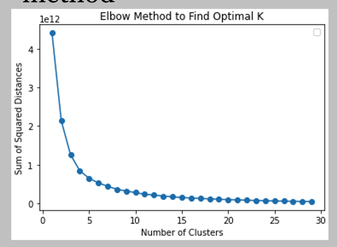
- Process historical data to look for associations between metadata and popularity
- 2 dense layers
- 9 months of dCache accesses
- ~160,000 records, 5 features

Features

- Present accesses
- Number of files in dataset
- Dataset size
- Bytes read from dataset
- Dataset type (experiment data, log, simulation, etc.)

K-means Clustering

- Assign datasets to groups based on their future popularity
- Present accesses vs. future accesses
- k=5, determined using elbow method



File Predictors

- Given a file access sequence, predicts future accesses

Algorithms

- First Successor (FS)
- Last Successor (LS)
- Last N Successors (LSN)
- Stable Successor (SS)
- First Stable Successor (FSS)
- Recent Popularity (RP)
- Predecessor Position (PP)
- Backup Predictor (BP)

RESULTS

- On the right are plots of the MAE loss and predicted vs. actual future access values. The test and train data are shown in both plots.
- Below is the results of clustering with k=5. Note the size and diagonal placement of the groups.
- The bottom right plot shows the performance of the file prediction algorithms. BP and RP perform best.

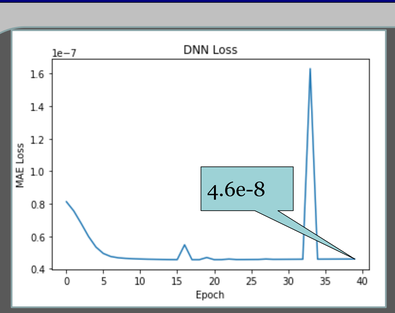


Figure 1: DNN loss during training

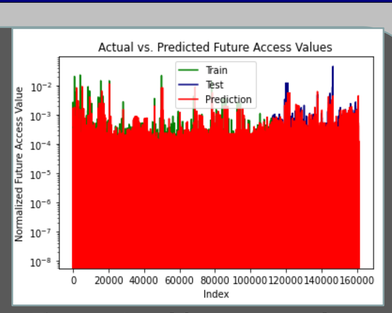


Figure 2: Actual future access values vs predicted future access values

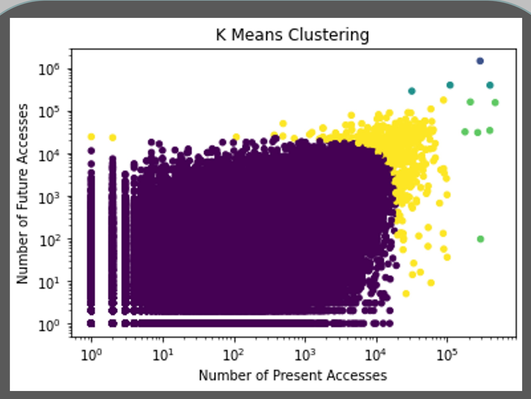


Figure 3: K-means clustering. K=5

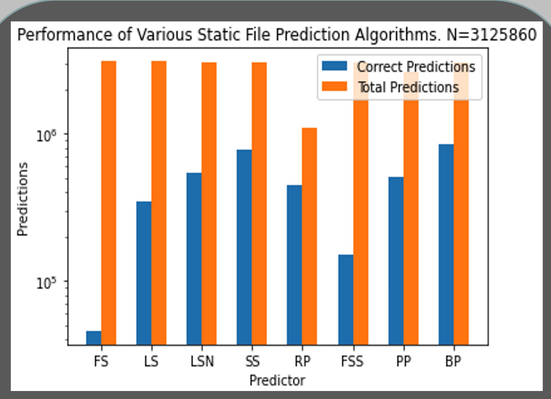


Figure 4: Number of correct and number of total predictions of each file prediction algorithm

CONCLUSION

- The deep neural network can accurately forecast how many times a dataset will be accessed the next day
- Clustering reveals that most datasets have fewer than 10000 future accesses, and those that have more than 10000 future accesses have many present accesses
- Backup Predictor using Recent Popularity and Last N Successors makes the highest raw number of correct predictions, at ~27.2%, Recent Popularity gets ~40.08% of the predictions it makes correct. These numbers are substantial compared to those produced by other predictors
- Generally, each method indicates that future dataset popularity can be predicted based on historical dataset information
- Future work will develop, simulate, and compare cache policies based on each method presented in this work

FURTHER READING

ACKNOWLEDGMENTS

Thanks to my mentors and collaborators, Alex Sim, John Wu, Shinjae Yoo, Hiro Ito, Eric Lancon and Vincent Garonne. This work was supported in part 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.