Predicting Dataset Popularity for Improved Distributed Content Caching in High Energy Physics

Malavikha Sudarshan[1], Alex Sim (advisor)[2], K. John Wu (advisor)[2]

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

ABSTRACT

In High Energy Physics (HEP), large-scale experiments generate massive amounts of data that are distributed globally. To reduce redundant data transfers and improve analysis efficiency, a disk caching system named XCache is used to manage data accesses. By analyzing 11 months of access logs (4.5 million requests), we identified patterns in dataset usage and developed a predictive model to forecast the popularity of frequently accessed datasets. Based on extensive exploratory data analysis, we found that pinging the most popular datasets could significantly improve access efficiency. To implement this pinging strategy, we attempted to predict which datasets would be the most popular in the near future.

BACKGROUND INFO

- The HEP community increasingly uses XCache for managing the disk caches for accessing the globally shared datasets.
- Data from the Southern California Petabyte Scale Cache (SoCal Repo) consisting of 23 nodes from CalTech, UCSD and ESnet at Sunnyvale.
- Study period: 09/2022 - 07/2023.
- ~ 4.5 million records: 3.7 million of these are cache hits, and some of the remaining 0.8 million cache misses may include re-transferred data.
- Our prediction model was ran on the top 25 most accessed datasets to assess its performance on dataset classes of varying popularity. This ranged from datasets with 310,337 accesses in the total time period to 24,421 accesses.

RESEARCH QUESTION

Can we reliably predict the future popularity of a dataset?

ANALYSIS AND PREDICTIONS

(1)

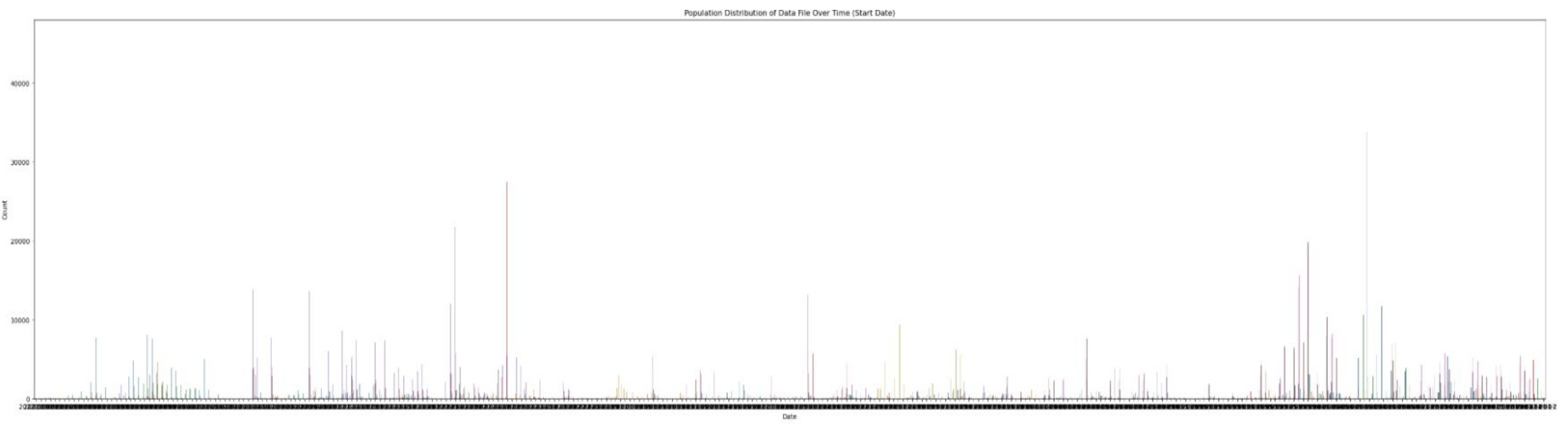


Figure 1: Population Distribution of Data File Type over Time (Access Start Date) across all 11 Months. Data accesses occur in bursts, as seen by the large spikes, and are difficult to predict.

(2)

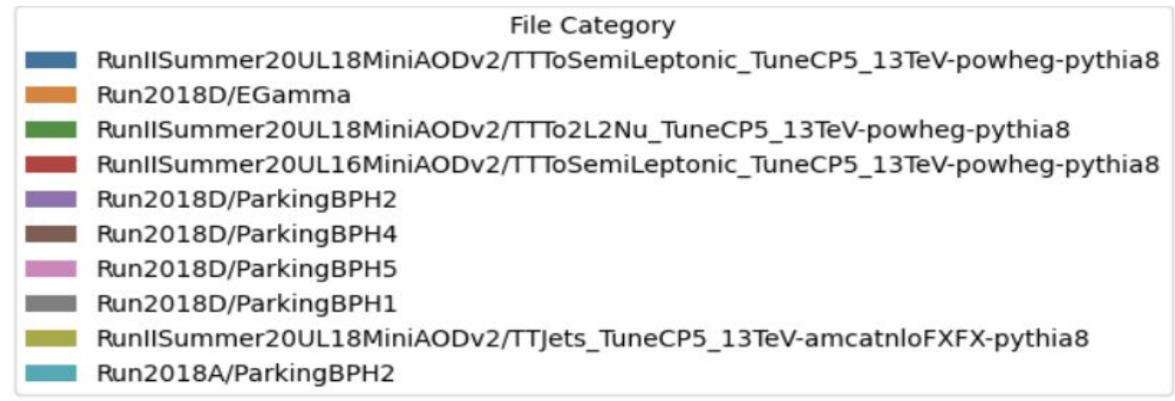


Figure 2: Legend. The dataset was split into “classes” using file name metadata.

(3)

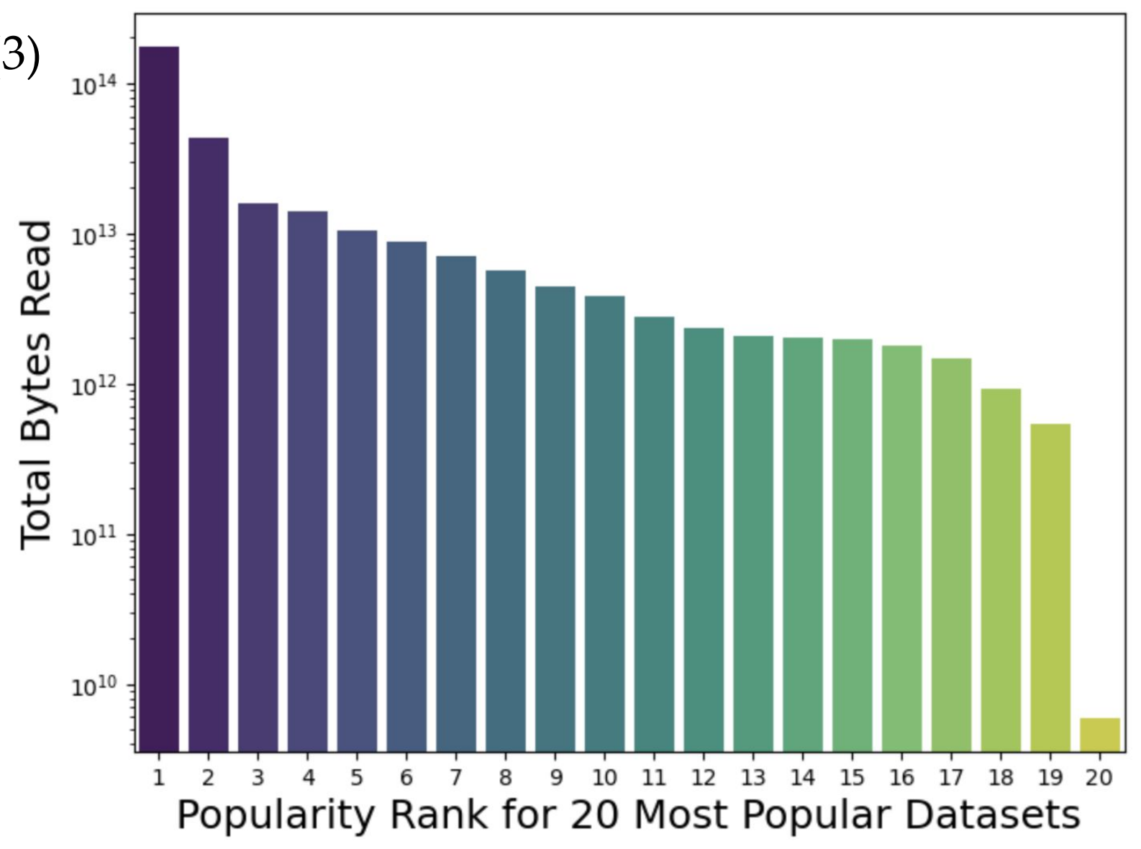
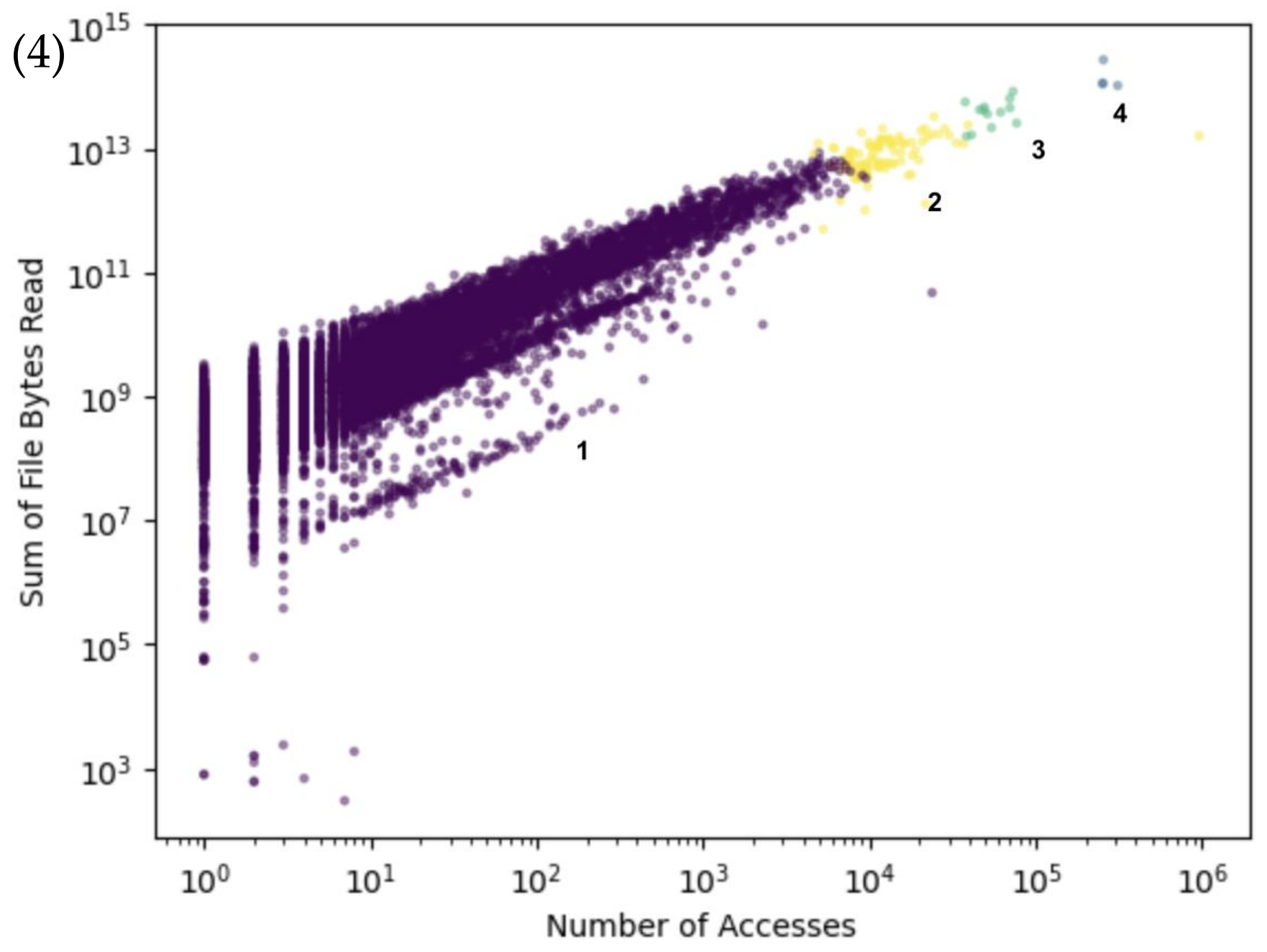


Figure 3: File Bytes Read in the Top 20 Datasets - even the 1st most popular dataset is much more frequently accessed than the next 19 combined, and the distribution of bytes from accessed files is highly skewed.

Figure 4: K-Means Clustering of All Datasets over the full 11-month Period, K = 4. As a few popular datasets account for the majority of bytes accessed, we propose pinging the most popular datasets in the disk cache to simplify cache management.

(4)



(5a) (5b)

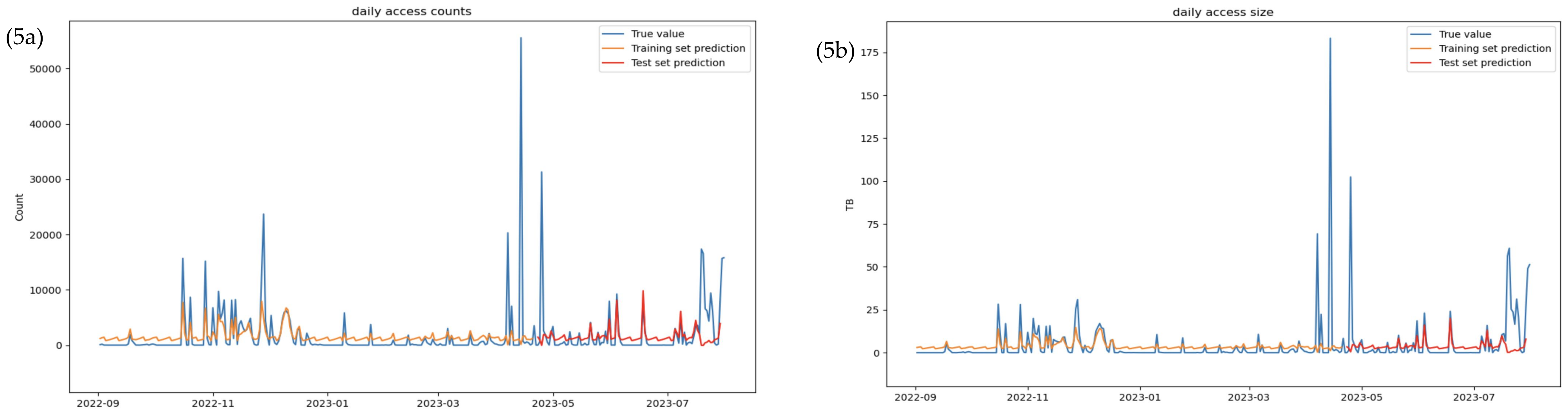


Figure 5: File Access Predictions for the Run2018D/ParkingBPH2, the most popular file using an LSTM model to predict accesses from September 2022 to August 2023: (a) daily access counts (b) daily access size. The time series is bursty with no apparent pattern in the intervals between bursts, making predictions difficult, but the LSTM does a decent job as the red/orange lines (predictions) follow the blue lines (actual values) fairly closely.

(6)

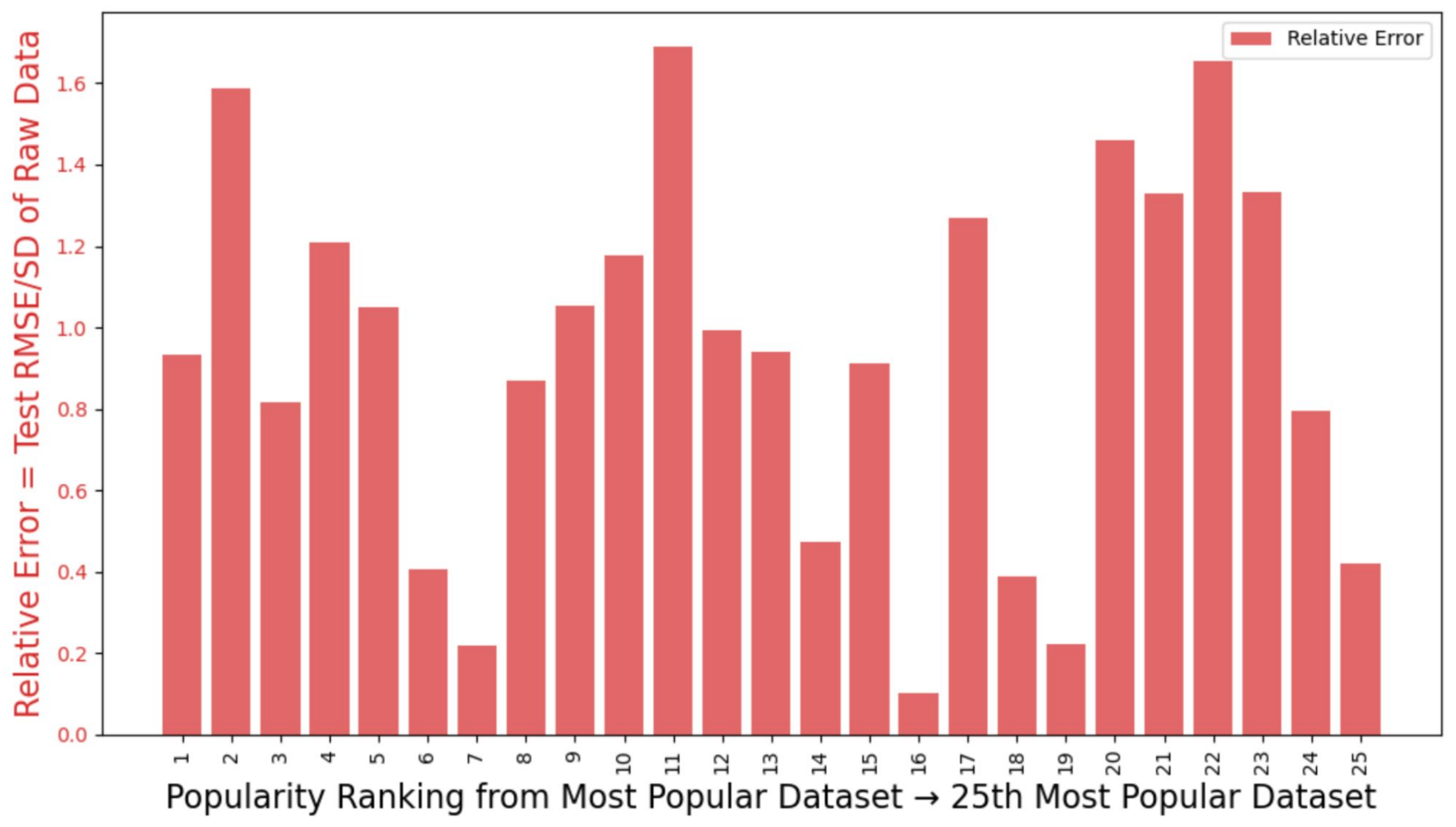


Figure 6: Relative Error values (Relative Error = Test RMSE/Standard Deviation (SD) of Raw Data) for the Top 25 Most Popular Datasets. The prediction errors are small (Relative RMSE values < 1 are considered to be small). Most of the more popular datasets have Relative Error values less than 1.

DISCUSSION

- An LSTM (Long Short-Term Memory) Neural Network model was employed for predicting dataset popularity, focusing on metrics such as file accesses, cache hits, and cache misses.
- The accuracy of the model was evaluated using RMSE (Root Mean Squared Error) values for both the training and test sets.
- Our LSTM model demonstrates strong performance in capturing general access trends, with a low mean relative error of 0.779 on both training and test datasets.
- Despite the model's effectiveness at predicting overall patterns, it struggles with anomalous access peaks, likely due to the inherent sparsity and variability in the data.
- To address these challenges, future work will focus on incorporating anomaly detection techniques to improve the model's robustness in handling irregular data patterns.

CONCLUSION

Our findings highlight the significant potential of LSTM-based models for optimizing data caching policies and improving the efficient utilization of file storage systems.

ACKNOWLEDGMENTS

Thanks to my mentors and collaborators, Alex Sim and John Wu. his 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. This research used resources of the National Energy Research Scientific Computing Center.