Grenze International Journal of Engineering and Technology, June Issue

Data Deduplication using Machine Learning

Narasimhula Revanth[1], Ravuri Varun[2], Ebbili Harshitha[3], J. Jane Rubel Angelina[4] and S.J.Subhashini[5]
[1-3]UG Students, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Virudhnagar, Tamil Nadu, India
Email: revanth.n1002@gmail.com, ravurivarun1@gmail.com, eharshitha.03@gmail.com
[4]Associate Professors, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Virudhnagar, Tamil Nadu, India
Email: j.janerubelangelina@klu.ac.in
[5]Department of IT, SRM Madurai College for Engineering and Technology Madurai,Tamil Nadu, India
Email: sjsubhashini1472@gmail.com

***Abstract*— Reducing data in storage systems is becoming more crucial as a practical way to cut down on data Center management expenses. Current post-deduplication delta-compression techniques combine delta compression with lossless compression and conventional data deduplication to optimize data reduction efficiency. Regretfully,we observe that because of their poor accuracy in identifying similar data blocks, current techniques achieve noticeably lower data-reduction ratios than the optimal.**

Data deduplication is a computing technique that removes duplicate copies of data that repeats. A similar and almost interchangeable term is single-instance data storage. This method will be used to reduce bytes that need to be sent during network data transfers same as to increase storage utilization. During the de-duplication process, distinct data segments, also known as byte patterns, are found and saved after analysis. Further chunks are compare with the stored copy as the analysis progresses, and if a match is found, the redundant chunk is substituted with a brief reference pointing to the stored chunk.

Reduced data storage or transfer requirements result from the possibility same as byte pattern occurring hundreds, thousands, or even more times (match frequency depends on chunk size). Comparing data segments to find duplicates is how most recognized methods for data de-duplication is implemented.

***Index Terms*— deduplication, spaCy , machine learning, VGG 16.**

I. INTRODUCTION

Data deduplication play a vital role in managing and optimizing storage resources by identifying and eliminating redundant data. Conventional deduplication methods often rely on byte sequence comparison algorithms, predefined patterns, or hash functions. However, with the advent machine learning technique (ML), there has been a paradigm shift towards more sophisticated and dynamic approaches to data deduplication. Machine learning makes data deduplication even more advanced by enabling systems to recognize patterns, correlations, and anomalies in datasets. Specifically helpful for managing a range of dynamic data structures is this evolution. Below summary of how machine learning is transforming the data deduplication market.

ML models may eventually adapt to shifting data patterns, changing the accuracy the deduplication process. This flexibility helps the system handle complex and dynamic data more effectively than rule-based approaches.

Grenze ID: 01.GIJET.10.2.62

Machine learning (ML) algorithm have the capability to decrease false positives by improving their understanding of the context and nuances of data. This is particularly important when strict deduplication might unintentionally result in data loss.

In conclusion, adding machine learning (ML) techniques to data deduplication processes adds a advanced intelligence and flexibility. By leveraging capability of algorithms from data, organizations can achieve more accurate and efficient deduplication, thereby optimizing storage resources and improving overall data management practices. Improving the deduplication process's accuracy, efficiency, and adaptability are the main goals of employing machine learning (ML) for data deduplication. These are the main goals:

- To provide reasonable throughput.
- Reduce Storage Costs.
- To prevent unnecessary file upload on the storage server that is already present.
- To reduces storage space requirements

II. Literature Survey

This study introduces the concept of a near duplicate dataset, or a quasi-duplicate version of a dataset. This release contains an unknown number of schema and instance adjustments (row and column insertions and deletions). This is an interesting concept for quality assurance, data integration, and research. To formalise these insertions and deletions, two parameters are introduced. Our methodology, which is on bases feature extraction and machine learning(ML), finds these almost-duplicate datasets. This different method from others methods in that it compares metadata vectors, which are summaries of the datasets, as opposed to comparing columns as in conventional methods. any information is available at a click. So, this helps a lot. To train these algorithms, we present a method to artificially generate training data. We run a number of experiments to measure the metric of different classifiers and determine the best parameters to use when creating training data. less research has done in this paper to generate more different cases to make learning more resilient. In future work, we will attempt to remove extra or unwanted features using feature selection algorithms.[1]

In this paper the analysis was done on medical dataset.Duplication which well known issue in bioinformatics that occurs when two or more database records corresponded to similar biological entity. Nevertheless, there are less methods for finding bioinformatic duplicates .Only a small number of machine learning(ML) have been proposed for bioinformatics, despite the reason they are used for finding similar data in general databases. We have tested one such approach on a set of similar data in nucleotide dataset which was labeled submitter. The study's best rule could only identify 0.2% of the duplicates, according to the final observation, and the final performance of all rules was dreadful. The method used reflects here that it can’t be Normalised to other data collections and has severe limitations.[2]

In the other paper storage space used before deduplication never changes, the larger the deduplication ratio, the more effective the duplication was to reduce final storage space consumption. In the rest of this paper, we propose Segment Deduplication, a solution to effective chunk-based file system deduplication using reinforcement learning.The system could perform differently in regard to deduplication ratio, depending on the dataset used, has few database may have significantly less duplicated data than the others.[3]

A reference for fingerprint data is used in this research to perform the deduplication process. Data transfer latency for remote systems and storage space reduction due to the removal of duplicate data are the two main advantages of the data deduplication technique thus far. An established problem in chunk-based deduplication is creating an effective fingerprint indexing to assist in locating duplicate data chunks. Keeping such a big index in RAM is not practicable, and large-scale storage cannot tolerate the slowness of a disk-based index with one seek for each incoming chunk. Such well-known difficulty is known as the chunk-lookup disk bottleneck problem. A small proportion of duplicate chunks are missed because it doesn't look for uncached fingerprints on disk. For this reason, making this sacrifice is necessary to improve the deduplication performance.[I4]

In this paper reducing data in storage repository have became most crucial as a practical way to cut down on data centre management expenses. Current post-deduplication delta-compression techniques combine delta compression with lossless compression and conventional data deduplication to optimize data reduction efficiency. Unfortunately, because of there limited accuracy in identifying similar data blocks, we observe that existing techniques achieve significantly lower data-reduction ratios than optimal .By improving reference search accuracy in post-deduplication delta compression and narrowing the variation between best and current data-reduction methods, we hope to increase a storage system's space efficiency. The current version of Deep Sketch requires a powerful GPU for DNN inference/training and thus introduces non-trivial performance and power overheads.[5]

III. EXISTING SURVEY

There are so many deduplication methods and algorithms existing before and now which remove the duplicate data from their storage systems to better performance and to overcome the overhead. Many systems like EMC Data Domain, Netapp deduplication and commvault especially in the storage industry have some complexities in their systems due to which they are less efficient in atmost functioning of their systems. The systems which are present cannot handle large datasets some cannot handle deduplication across heterogeneous storage systems. Other systems require significant computational resources for inline deduplication processes. Initial backup of deduplication systems maybe delayed when dealed with larger datasets. Considering all these weaker sections improvements can be made to make efficient powerful systems like integrating the systems with machine learning (ML) method which can enhance deduplication efficiency by predicting data patterns, we can also combine deduplication with latest compression techniques for storage efficiency, innovations in deduplication algorithms and methodologies can minimize the impact on backup and recovery processes.

IV. PROPOSED SYSTEM

Designing an effective and efficient data deduplication system involves addressing the drawbacks of existing techniques while leveraging improvements in technology.

A. Hybrid Deduplication approach

Combining both inline and and post-process deduplication strategies. Utilizing the advantages of inline deduplication for real-time data reduction and post-process deduplication for reducing the impact on prefect accuracy.

B. Machine Learning Integration

We are implementing machine learning (ML) algorithm to predict the patterns. It can adapt to changing the data patterns and improving accuracy overtime.

C. Enhanced Compression Techniques

We shall integrate advanced compression techniques like spacy module used in NLP (stemming and lemmatization) for deduplicated text removal which further reduces storage space requirements.

D. Continuous monitoring and Optimization

Implementing continuous monitoring of deduplication performance and enable automatic adjustments and optimizations based on changing work loads and data patterns.
Firstly we create a user interface using python gui where a user can login using the username and password then they will be able to upload text files and images within a folder.After uploading the system eliminates the repeated text in the text files,in an excel etc and repeated images within a file can be deleted using machine learning algorithm.

V. METHODOLOGY

For eliminating duplicated text we are achieving this with the help of Natural language processing (NLP) techniques. One such method is spaCy it is an open-source natural language processing (NLP) library in Python which is mostly used for various NLP tasks, including lemmatization and stop word removal.
spaCy uses pre-trained models for various languages that include lemmatization information. It determines the correct lemma by examining the text of each word in a phrase. we use spaCy NLP technique for finding text duplication here we give input and apply the algorithm then it checks the lemma and stop words and remove all the similar kind odspaCy provides a list of default stop words for each language. During text processing, these stop words can be filtered out.spaCy is designed for efficiency, and its models are optimized for speed and memory usage. This makes it suitable for large-scale processing tasks. spaCy provides pre-trained models for various languages, allowing you to perform lemmatization and stop word removal in different languages.

VGG16

Here we take the input from the user and we classify the data into type of data VGG16 is used for image type of data here we take different images and move through VGG16 CNN model then is passes through 16 layers of it and checks for the duplicate. If we found any duplicate then we store the first image in the next occurring images for this pooling technique is used from that we find the similarities. A 16-layer convolutional neural network(CNN) is called VGG-16. A pre trained version of the network, trained on over a ten lakh images from the Image-Net

dataset. Images of thousand different objects are categorized, including a keyboard, mouse, pencil, and numerous animals, can be identified by the pretrained network.With 92.7% accuracy, the object detection and classification algorithm VGG16 can classify thousand images into thousand distinct categories. It's well-liked image classification algorithm that is simple to use with transfer learning.

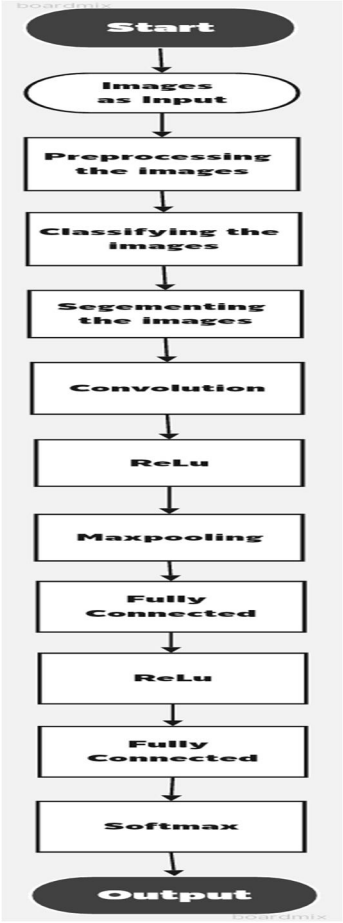

Figure 1.VGG16 Flow Diagram.

The sixteen in VGG16 stands for sixteen weighted layers. Although VGG16 contains twenty-one layers total—sixteen (CNN) layers, with 5 Max Pooling layers, and three Dense layers—it only has sixteen weight layers, or learnable parameters layers.VGG16 accepts a 224, 244 input tensor size with three RGB channels. The most attractive element of VGG16 is that its convolution layers of a 3x3 filter with stride 1 are its importance, without having a lot of hyper-parameters. The padding and maxpool layer of a 2x2 filter with stride 2 are also always used. The architecture as a whole has the convolution and max pool layers placed in the same order. There are 64 filters in Conv-1 Layer, 128 filters in Conv-2, 256 filters in Conv-3, and 512 filters in Conv-4 and Conv-5.Three Fully-Connected (FC) layers are stacked after a stack of convolutional layers. The first two have 4096 channels each, while the third classifies thousands of ways using ILSVRC, yielding thousand channels (one for each class). The final layer is called Soft-max.

VI. Results And Discussion

This survey study presents the most recent findings on deduplication techniques and is an addition to the previous surveys. There have been attempts to address the unresolved research challenges in deduplication algorithms in large distributed storage systems.Our system eliminates the deduplicated data from various forms of data like text images, excel files etc. Machine learning (ML) has been combined with deduplication system along with advanced compression techniques to provide better efficiency for removing duplicate text files NLP techniques were used for running accurately. And a neural network (NN) model was used for removing duplicate images within a file.

VII. SAMPLE SCREENSHOT

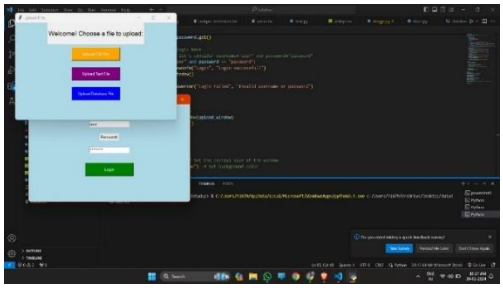

Figure 2. Sample Screenshot 1

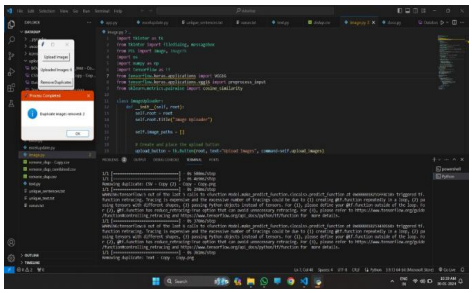

Figure 3. Sample Screenshot 2.

VIII. CONCLUSION

In the current environment, data deduplication and cloud computing seems being hottest trends. Data deduplication is becoming most demand as cloud computing continues to grow and requires digital data being stored efficiently in huge storage systems. Massive data loads on storage systems are highlighting the need to concentrate on learning new methods for deleting redundant data. The opportunity for eliminating redundant data exists in cloud computing due to the advancements made in data deduplication and its diverse methods. Deduplication is a crucial method for cutting back on bandwidth, energy use, and storage costs.

This study provided a thorough analysis of data deduplication methods. This survey examines and evaluates deduplication methods in great detail. We have also thoroughly examined a few recent surveys covering related subjects. Previous studies in this area have concentrated on storage-type-based deduplication methods. This survey's main goal is to investigate text- and multimedia-based deduplication methods. The results suggest that there are numerous issues with deduplication, which needed to resolved in text- and multimedia-based deduplication.

REFERENCES

[1] Marc Chevallier, Nicoleta Rogovschi, Faouzi Boufarès, Nistor Grozavu, Charly Clairmont. Detecting Near Duplicate Dataset with Machine Learning. International Journal of Computer Information Systems and Industrial Management Applications, In press.

[2] Chen, Qingyu, Justin Zobel, and Karin Verspoor. "Evaluation of a machine learning duplicate detection method for bioinformatics databases." Proceedings of the ACM Ninth International Workshop on Data and Text Mining in Biomedical Informatics. 2015.

[3] Yuan, Xincheng, "Whole File Chunk Based Deduplication Using Reinforcement Learning" (2022). Master's Projects. 1080.

[4] Xu, Guangping, et al. "Lipa: A learning-based indexing and prefetching approach for data deduplication." 2019 35th Symposium on mass storage systems and technologies (MSST). IEEE, 2019.

[5] Park, Jisung, et al. "{DeepSketch}: A new machine {Learning-Based} reference search technique for {Post-Deduplication} delta compression." 20th USENIX Conference on File and Storage Technologies (FAST 22). 2022.

[6] Zhao, Mark, et al. "RecD: Deduplication for End-to-End Deep Learning Recommendation Model Training Infrastructure." Proceedings of Machine Learning and Systems 5 (2023).

[7] Mageshkumar, N., and L. Lakshmanan. "Intelligent data deduplication with Deep Transfer Learning Enabled Classification Model for Cloud-based Healthcare System." Expert Systems with Applications 215 (2023): 119257.

[8] Adimoolam, Yeshwanth Kumar, et al. "Efficient Deduplication and Leakage Detection in Large Scale Image Datasets with a focus on the CrowdAI Mapping Challenge Dataset." arXiv preprint arXiv:2304.02296 (2023).

[9] Lin, Lifang, et al. "Inde: An inline data deduplication approach via adaptive detection of valid container utilization." ACM Transactions on Storage 19.1 (2023): 1-27.

[10] Boinski, Pawel, et al. "On evaluating text similarity measures for customer data deduplication." Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing. 2023.