Chunk Content Is Not Enough:
Chunk-Context Aware Resemblance Detection For Deduplication Delta Compression

Xuming Ye, Xiaoye Xue, Wenlong Tian, Zhiyong Xu, Weijun Xiao, Ruixuan Li, Yaping Wan
Outline

• Background And Motivation
• Traditional solutions
• Problem & Solution & Design
• Evaluation
• Conclusion
1. Background And Motivation
Nowadays, data deduplication is critical in the storage system. With the increase of the devices, such as IoT devices, mobile device, etc..., the ever-increasing demand of storage space is pressing.
Data deduplication

Redundancy Deduplication
extra copies of the same data are deleted, leaving only one copy to be stored.

Resemblance Deduplication
Although redundant data elimination is efficient, but in the storage system, there are also much similar data.

2. Traditional solutions
Figure 1: An example of delta compression on two similar chunks with the three typical steps: ① computing similarity, ② indexing, and ③ delta encoding.
3. Problem & Solution & Design
Workflow

Training Process

1. Splitting the data into chunks
2. N-sub-chunk shingles based initial features extraction
3. Embed the chunk-context into the chunk feature

Predicting Process

1. Splitting the data into chunks
2. N-sub-chunk shingles based initial features extraction
3. Predict the chunk-context aware feature
4. Delta encoding based on context-aware features
CARD

- Extract Features

N-sub-chunk scheme $c_k_i$

\[ v(f) = \frac{1}{n} \sum_{i=0}^{n} f_i \]

Dimension=$D$

- $f_1$: $\text{vector}_{f_1}$
- $f_2$: $\text{vector}_{f_2}$
- $f_3$: $\text{vector}_{f_3}$
- $f_4$: $\text{vector}_{f_4}$
- $f_{n-1}$: $\text{vector}_{f_{n-1}}$
- $f_n$: $\text{vector}_{f_n}$

Initial vector $\text{vector}_i$
Workflow

1. Splitting the data into chunks
2. N-sub-chunk shingles based initial features extraction
3. Embed the chunk-context into the chunk feature

Training Process:
- Input bits stream
- Split into chunks
- N-sub-chunk shingles
- Embedding chunks
- Contextual vector

Predicting Process:
- Input bits stream
- Split into chunks
- N-sub-chunk shingles
- Embedding chunks
- Contextual vector
- Delta encoding based on context-aware features

10/15
- Extract Features

BP-Neural Network-based Chunk-Context Aware Model
Workflow

Training Process

1. Splitting the data into chunks
2. N-sub-chunk shingles based initial features extraction
3. Embed the chunk-context into the chunk feature

Predicting Process

1. Splitting the data into chunks
2. N-sub-chunk shingles based initial features extraction
3. Predict the chunk-context aware feature
4. Delta encoding based on context-aware features
4. Evaluation
<table>
<thead>
<tr>
<th>Name</th>
<th>Workload descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAR</td>
<td>Versions of Linux kernel source code. Each version is packaged as a tar file.</td>
</tr>
<tr>
<td>BDB</td>
<td>Backups of the real company database.</td>
</tr>
<tr>
<td>VMDK</td>
<td>Snapshots of an Ubuntu 18.04 VM image.</td>
</tr>
</tbody>
</table>
5. Conclusion
Conclusion

According to the presentation above, our advantages as the following.

- Features to vector
- N-sub-chunk shingle scheme
- Chunk-Context Model
CARD

Thanks for watching

our arxiv: https://arxiv.org/abs/2106.01273