Updatable Learned Index with Precise Positions
Experiment

Wu J, Zhang Y, Chen S, et al., **2021 VLDB**

2024. 02. 21

Presentation by Nakyeong Kim, Suhwan Shin
nkkim@dankook.ac.kr, shshin@dankook.ac.kr

Contents

1. Previous Experiment

Metric for Evaluating Model

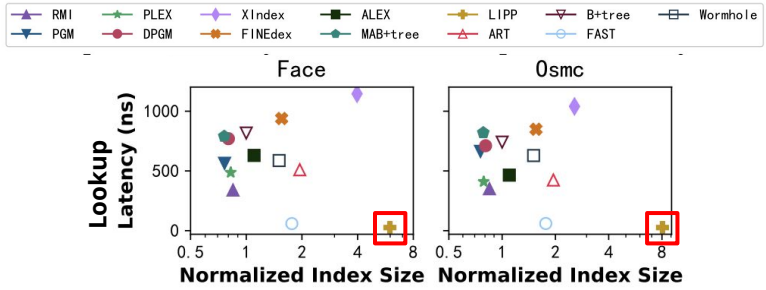
Most nodes have less than 30% of all entries

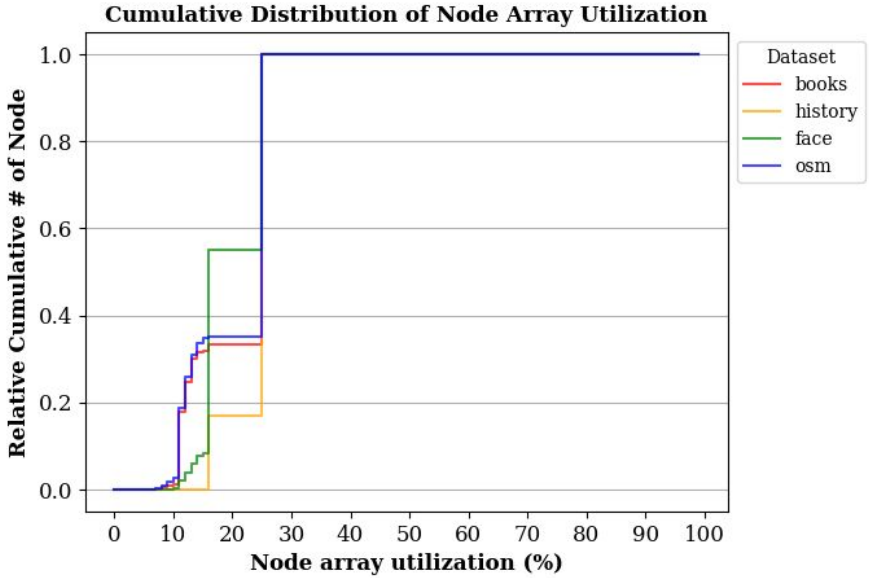
→ There exists upper bound

$$T_{\mathcal{M}} = \max_{l \in [0, L-1]} |\{k \in \mathcal{K} | \mathcal{M}(k) == l\}|$$

We observe that there exists an upper bound for the minimum $T_{\mathcal{M}}$, i.e. $\exists \mathcal{M}, T_{\mathcal{M}} \leq \lceil \frac{N}{3} \rceil$ where N is the number of keys in $\mathcal{K}$, i.e. $N = |\mathcal{K}|$. However, the $\lceil \frac{N}{3} \rceil$ may not be the tightest upper bound in many cases. Thus, our goal is to find a best model $\mathcal{M} = A\mathcal{G}(k) + b$ with the minimum conflict degree $T_{\mathcal{M}}$.

Any consecutive *T+1* elements should not conflict in the same position by conflict degree *T*





1.1. Adjustment Strategy

When to Adjust

1. Insert key(s)

2. Update and check statistics of nodes in the traversal path

3. Trigger adjustment on a chosen node when certain conditions are satisfied
   a. $\frac{n.element\_num}{n.build\_num} \geq \beta$ β is set to 2 by default
   b. $\frac{n.conflict\_num}{n.element\_num - n.build\_num} \geq \alpha$ we set the threshold $\alpha = 0.1$

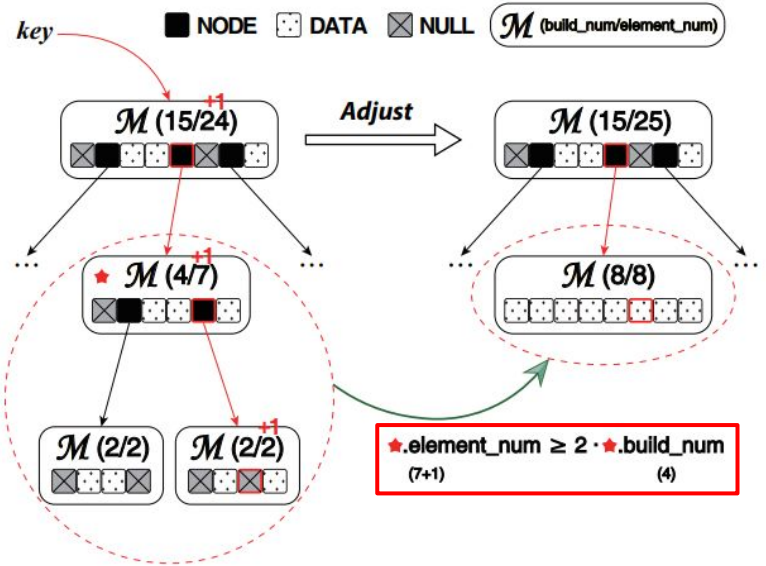


Figure 4: Node Adjustment

1.1. Adjustment Strategy

How to Adjust

1. Collect all elements(keys) in the subtree rooted at node by sequential traversal

2. Build a partial tree on the elements

3. Update the pointer of the original node to point of the new node(tree)

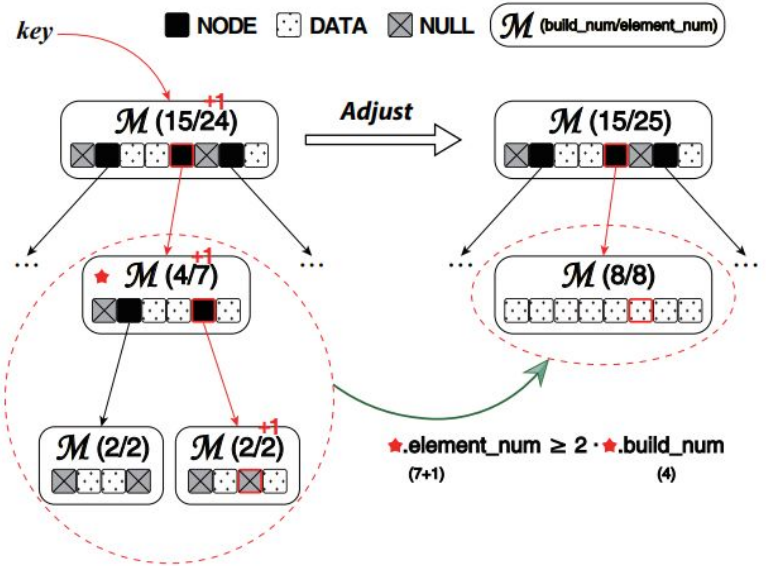
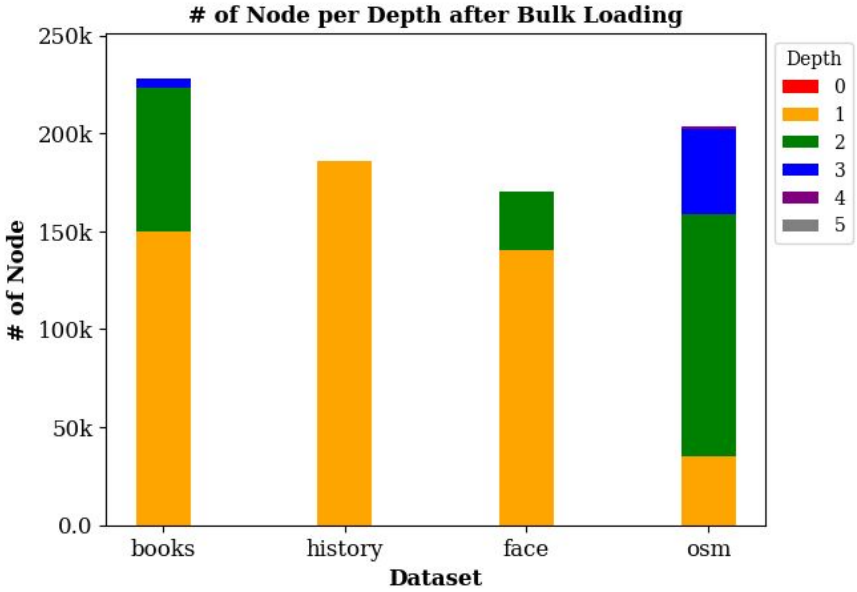
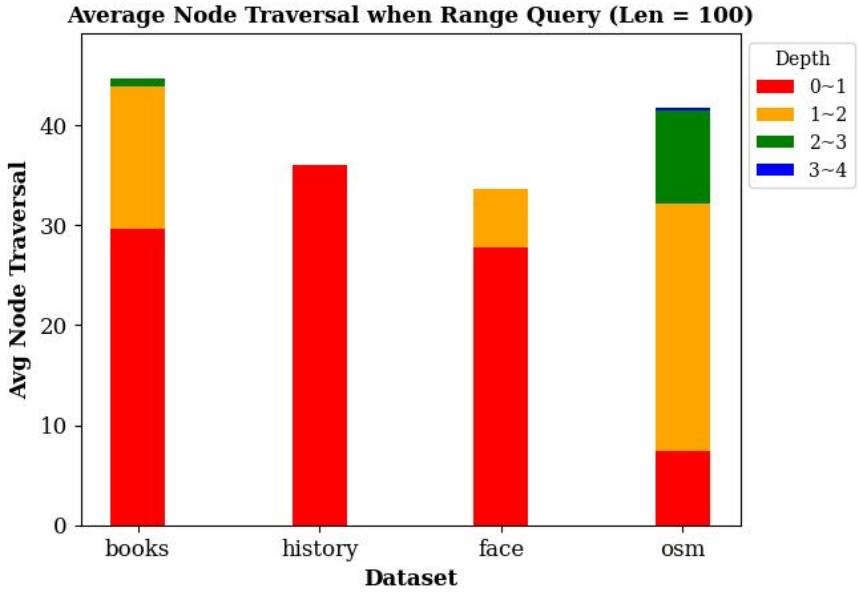


Figure 4: Node Adjustment

2.1. Experiment

Tendency by Dataset

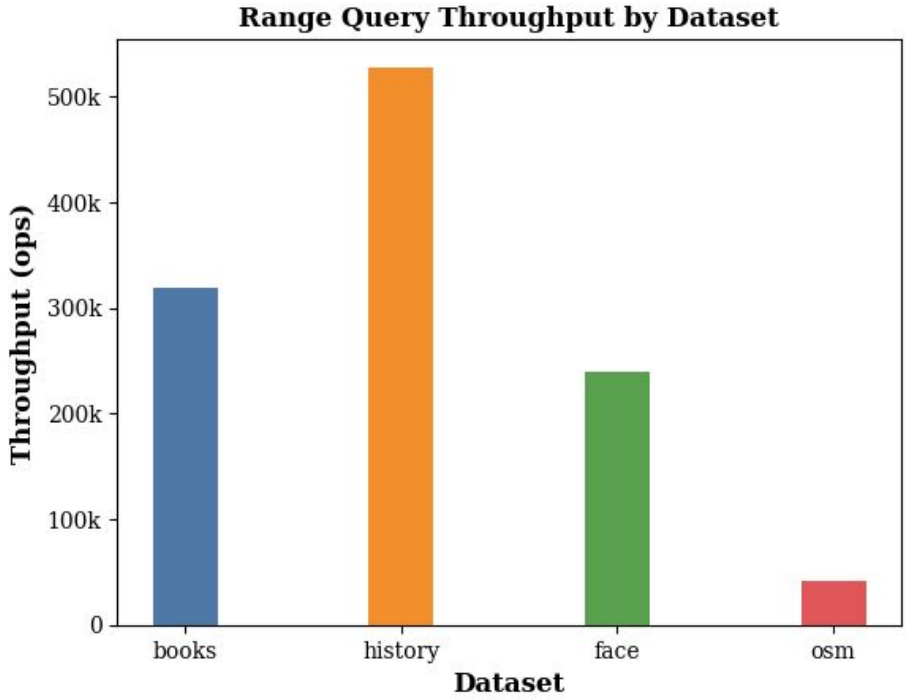




2.2.1. Experiment

scan_num: 100
iteration: 1M
table_size: 100M (key_range: 200M)
operations: range query only

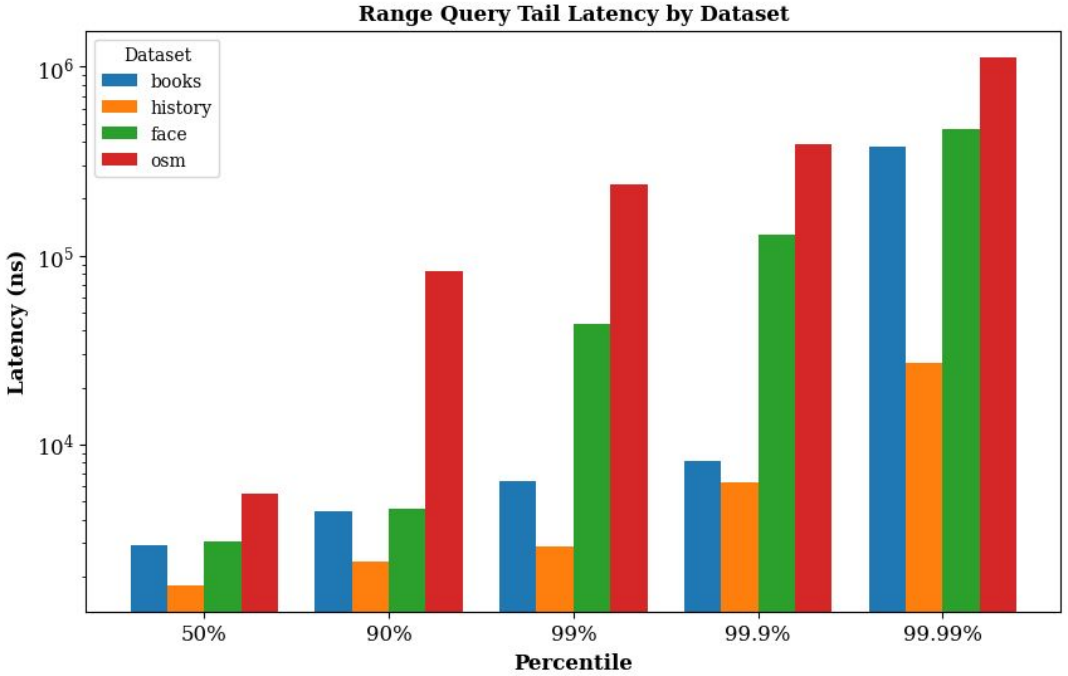
Range Query Throughput by Dataset



2.2.2. Experiment

scan_num: 100
iteration: 1M
table_size: 100M (key_range: 200M)
operations: range query only

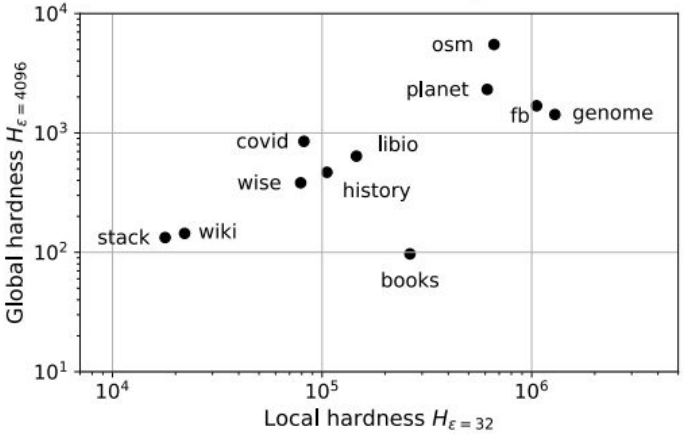
Range Query Latency by Dataset



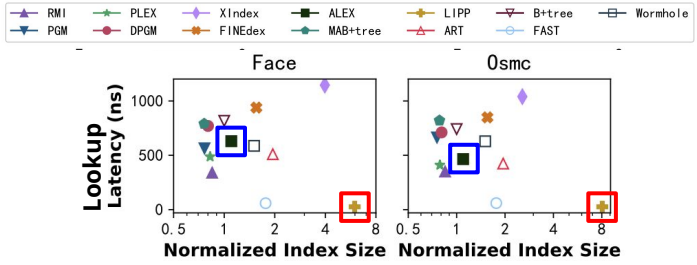
2.3. Dataset Hardness

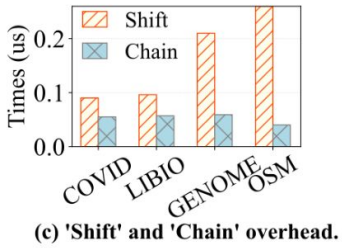
New Criteria

- Previous experiments defined data hardness as the number of segments (optimal PLA model)
- But our approach is based on **conflict count**, so it is not compatible
- Need to quantify to other criteria
- Plan to use definition of conflict degree of node



3. Next Step





(c) 'Shift' and 'Chain' overhead.

Conflict-Tolerance Strategy

- Focus on "**Conflict**" (SAF)

- Work on both factors at the same time
  - Motivation
    - Compare LIPP(Model+Chaining) and ALEX(Model+Shifting) with same index size
      - Conflicts
      - Shift/Chaining overhead

  - Design
    - Study meaning of TM's N/3 (What happened when more proportion)
    - Try shifting until some threshold (e.g., array size * *a*)
      After, use chaining

Q&A

Thank you!