Work Progress

kNN Search with Parallel Incremental Query Answering

Jaouhara Chanchaf

Friday Dec. 16th, 2022

1. Summary

Done:

AI 1	Masure the pruning power of PEXESO in different settings (change the parameters, the dataset size etc.)
AI 2	Validate hypothesis about pivot selection impacting the filtering power of PEXESO.

Not started: AI 3 - AI 9

2. Impact of τ (distance threshold) on Data Access

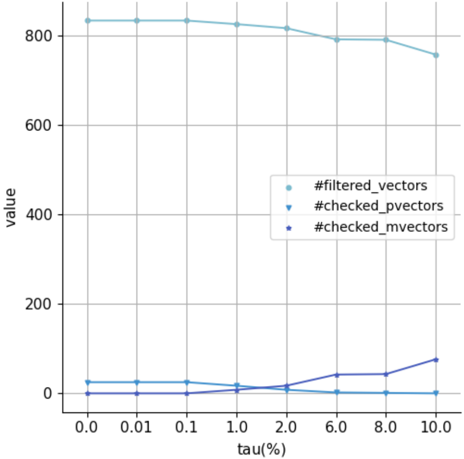

Figure 1: PEXESO: impact of τ on the number of filtered vectors (by lemmas 1, 3 and 4) and the number of checked vectors in the pivot space (pvectors) and the number of checked vectors in the metric space (mvectors)

- **Dataset:** 10 tables, 833 vectors.
- **Query:** 1 Query of size 1.
- **%Non empty leaf cells:** 4.9%.
- **Settings:** P = 3, m = 4, T = 0.6, fftscale = 30.

2. Impact of τ (distance threshold) on Data Access

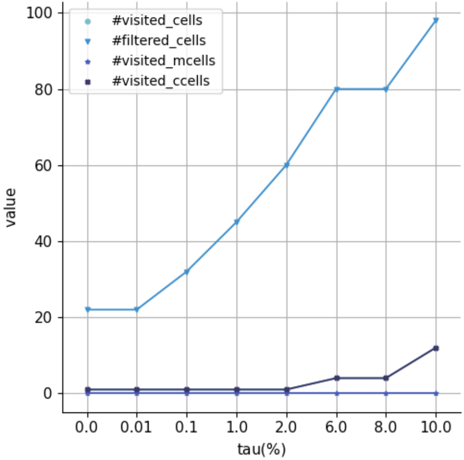

Figure 2: PEXESO: impact of τ on the number of filtered cells (by lemmas 3 and 4) and the number of visited matching cells (mcells) and the number of visited candidate cells (ccells).

- **Dataset:** 10 tables, 833 vectors.
- **Query:** 1 Query of size 1.
- **%Non empty leaf cells:** 4.9%.
- **Settings:** P = 2, m = 2, T = 0.6, fftscale = 30.

3. Impact of data distribution (in the pivot space) on Data Access

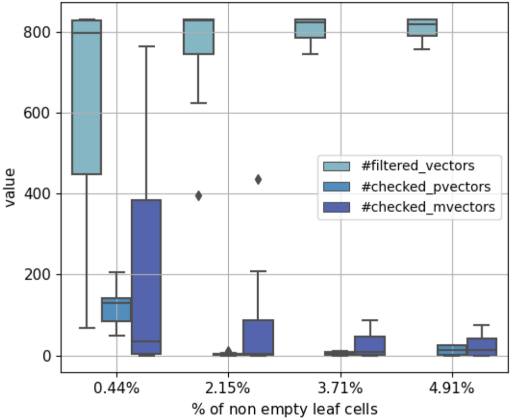

Figure 3: PEXESO: impact of the number of empty leaf cells on the number of filtered vectors (by lemmas 1, 3 and 4) and the number of checked vectors in the pivot space (pvectors) and the number of checked vectors in the metric space (mvectors).

- **Dataset:** 10 tables, 833 vectors.
- **Query:** 1 Query of size 1.

- **Settings:** P = 2, m = 2, T = 0.6, fftscale = {1, 30},
  $\tau = \{0\%, 0.01\%, 0.1\%, 0.2\%, 0.6\%, 0.8\%, 10\%\}$.

3. Impact of data distribution (in the pivot space) on Data Access

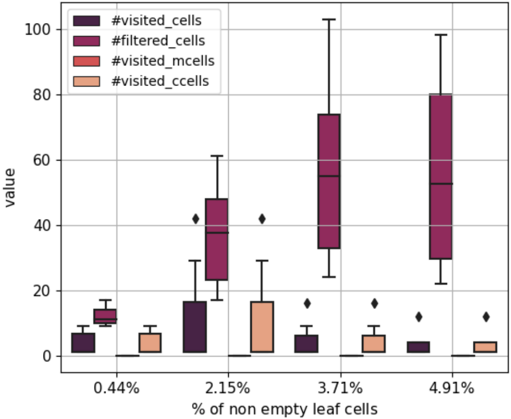

Figure 4: PEXESO: impact of the number of empty leaf cells on the number of filtered cells (by lemmas 3 and 4) and the number of visited matching cells (mcells) and the number of visited candidate cells (ccells).

- **Dataset:** 10 tables, 833 vectors.
- **Query:** 1 Query of size 1.

- **Settings:** P = 2, m = 2, T = 0.6, fftscale = 30, $\tau = \{0\%, 0.01\%, 0.1\%, 0.2\%, 0.6\%, 0.8\%, 10\%\}$.

4. Discussion

- To get the best data distribution (i.e maximum nb of non empty leaf cells) the pivot selection algorithm must be repeated several times! This task is time consuming especially when the dataset size is large.
- Small τ values yield better filtering Figure 1. A large τ values results in less filtering hence most data vectors are evaluated in the metric space.
- Increasing τ leads to an increase in the number of candidate cells Figure 2 and consequently increases the number of filtered cells.
- Ideally we want to represent high dimensional data vectors in a lower dimension space (pivot space) while maximizing the variance of the data along the dimensions of the pivots. This can be translated to the number of non empty leaf cells in PEXESO grid. The more non leaf cells we have the higher is the data variance.
- Figures 3 and 4 illustrate how the higher the data variance the more filtering PEXESO can perform.