Supplementary Material for “Reconfigurable Inverted Index”

Yusuke Matsui
National Institute of Informatics
matsui@nii.ac.jp

Ryota Hinami
The University of Tokyo
hinami@nii.ac.jp

Shin’ichi Satoh
National Institute of Informatics
satoh@nii.ac.jp

1 DECISION OF THE THRESHOLD θ

We show how to decide the threshold parameter θ. As discussed in Sec. 4.2, PQ-linear-scan is selected if the number of the target identifier ($|\mathcal{S}|$) is smaller than the threshold θ, otherwise inverted-index is selected. It is hard to decide the optimal θ preliminary because it depends on several parameters. Therefore, we simply run the both methods (PQ-linear-scan and inverted-index) several times when the data structure is built (i.e., when the reconfigure function is called). By fitting a 1D line for given observations, we empirically decide the best parameter. This works perfectly for all cases in our evaluation.

As shown in Fig. 1, we found that the θ depends on the candidate length L. With larger L, θ also becomes large. So we set L from five values: $L \in \{\sqrt{N}, 2\sqrt{N}, 4\sqrt{N}, 8\sqrt{N}, 16\sqrt{N}\}$. For each L, PQ-linear-scan and inverted-index are compared by drawing a graph like Fig. 1, then the best θ for each L is measured. This can be efficiently conducted by recursively selecting $|\mathcal{S}|$. The measured θ is plotted as shown in Fig. 2. As this figure shows, the results are roughly proportional to L. Thus we simply fit 1D line to these measured value, and obtain the line as the form of linear function:

$$\theta = \alpha_1 L + \alpha_2. \tag{1}$$

The two coefficients α_1, α_2 are stored. With this function, we can compute θ for any L.

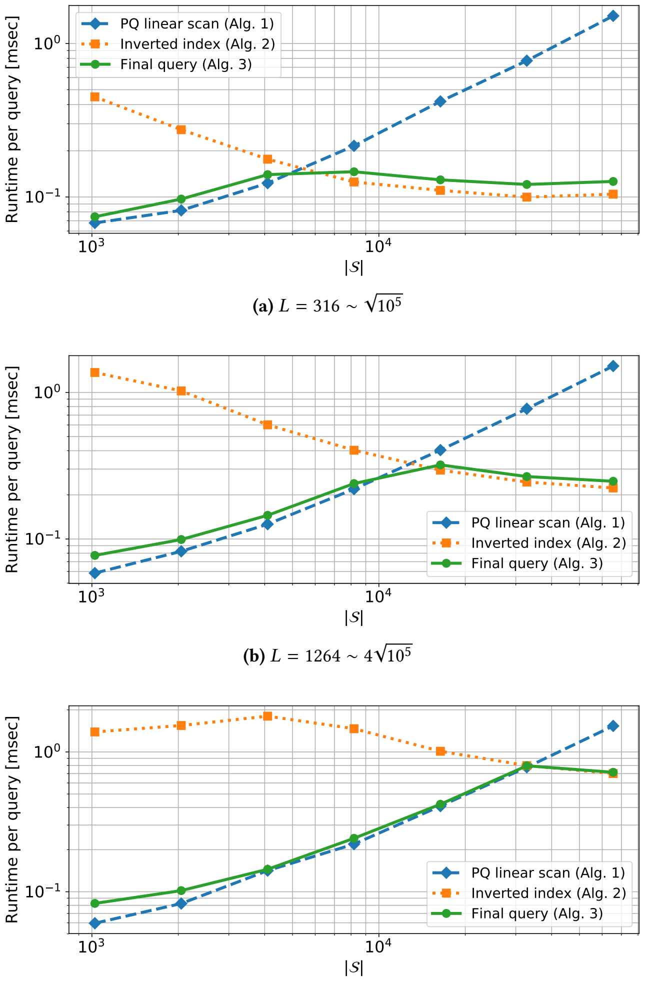
In the search phase given L, the threshold θ is decided using Eq. (1). This works perfectly as shown in the “Final query” in Fig. 1.

2 ASSUMPTION OF THE UNIFORM DISTRIBUTION

The computational complexities described in Table 1 of the main manuscript are based on the assumption that both the number of items per center and subset identifiers are equally distributed. Although not theoretically justified, this assumption is often reasonable for real world data.

Regarding the number of items per center, we run k-means-based clustering [1] for input vectors (Sec. 4.1 in the main manuscript). This means that the number of items is likely to be balanced due to the nature of k-means. Note that, of course, this is not always true because (1) we can come up with an extreme case where items must not be balanced such as all items are identical, or (2) the balanced results do not always be kept after new items are added.

In terms of subset identifiers, we believe our assumption of the equal distribution is a reasonable estimation. If items are not distributed equally, many items lie in the same posting list. This makes the search faster because we only need to consider a smaller number of posting lists.



(a) $L = 316 \sim \sqrt{10^5}$

(b) $L = 1264 \sim 4\sqrt{10^5}$

(c) $L = 5059 \sim 16\sqrt{10^5}$

Figure 1: Comparison of each methods with different $|\mathcal{S}|$. The results with different L are illustrated. Random valued 128-dimensional data ($N = 10^5, M = 32$) are used here.

3 RELATED WORK: BINARY HASHING

Note that as a complementary approach to PQ, there are many methods based on binary hashing [2, 3]. We do not focus on these, because their accuracy is usually lower than PQ-based methods with the same bit-length.

REFERENCES

[1] Yusuke Matsui, Keisuke Ogaki, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2017. PQk-means: Billion-scale Clustering for Product-quantized Codes. In *Proc. MM*.

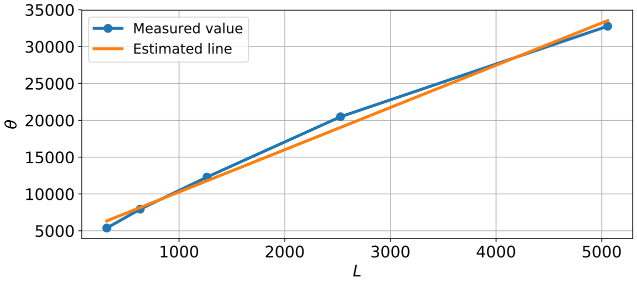


Figure 2: 1D line fitting over the observations.

[2] Jun Wang, Wei Liu, Sanjiv Kumar, and Shih-Fu Chang. 2016. Learning to Hash for Indexing Big Data - A Survey. *Proc. IEEE* 104, 1 (2016), 34–57.

[3] Jingdong Wang, Ting Zhang, Jingkuan Song, Nicu Sebe, and Heng Tao Shen. 2018. A Survey on Learning to Hash. *IEEE TPAMI* 40, 4 (2018), 769–790.