Qs Rank and efficient ϵ-neighbor search

Arthur Darcet & Yohann Salaun

March 26, 2013

1 Overview

The QsRank algorithm, described in [1], is a method that allows a ranking for binary hash codes to efficiently perform ϵ-neighbors search in large data. The aim of this problem is to, for a given query q and a given radius ϵ, quickly finds the data subset $Y = (y_i)_i$ such that $\forall i,\ ||y_i - q|| < \epsilon$.
Methods based on the Hamming distance have been recently developed to solve efficiently such a problem for large data. However these latter methods often lack of precision due to their binarization of the query which induces a lack of speed in the performances of their algorithm. The methods presented here, performs a more accurate ranking for hash codes and thus allows a faster and more efficient method for ϵ-neighbors retrieval.

2 Theoretical Description

The algorithm of ϵ-neighbor search using the QsRank works in two times:

1. Hash code generation of the data for faster retrieval
2. Ranking of the hash code depending on a query

The first part can be computed before any requests whereas the second needs the parameters of the search (query q and neighbor distance ϵ).

2.1 Hash Codes generation

Such search algorithm are used for huge data with components in high dimensional space. Thus, in order to generate simple hash codes for each input data point $x \in \mathbb{R}^d$, a Principal Component Analysis (PCA) is computed.
The hash code h of x is then computed by taking the sign of the PCA-projections:

$$\forall j \in [1; d],\ h_j = \begin{cases} 1 \text{ if } (PCA(x, \{x_i\}_i))_j > 0 \\ 0 \text{ if } (PCA(x, \{x_i\}_i))_j \leq 0 \end{cases}$$

In practice, only the first components of the hash codes are computed.
The benefits given by the PCA are numerous:

1. The PCA allows a dimension decrease but keeps most of the information. This way, hash codes are shorter and the retrieval is still efficient.
2. The PCA is an orthogonal projection that preserves the L^2-norm. Thus, the ϵ-ball around a query is still meaningful after PCA.
3. The PCA values of the input data points are uncorrelated which will lead to an efficient ranking with Qs Rank.

2.2 QsRank for Hash Codes ranking

We suppose, that the projection y of the input data points x are distributed along a probability distribution function p. Then, with a given query q, a neighbor distance ϵ and a hash code h, the Qs Rank formula is defined by:

$$\text{QsRank}(q, h, \epsilon) = \frac{\int_{NN(q,\epsilon)\cap S(h)} p(y)dy}{\int_{NN(q,\epsilon)} p(y)dy}$$

where:

- $NN(q, \epsilon) = \{y \in \mathbb{R}^d \text{ s.t. } ||y - q|| < \epsilon\}$ is the ϵ-ball around the query q
- $S(h) = \{y \in \mathbb{R}^d \text{ s.t. } \forall i \in [1; d] y_i h_i > 0\}$ is the set described by the hash code h

The QsRank can only be seen as a probability, with Bayes rule:

$$\text{QsRank}(q, h, \epsilon) = \frac{\mathbb{P}(y \in NN(q, \epsilon) \cap S(h))}{\mathbb{P}(y \in NN(q, \epsilon))} = \mathbb{P}(y \in S(h) | y \in NN(q, \epsilon))$$

Thus, the QsRank only ranks hash codes with respect to their probability of containing many ϵ-neighbors.

2.3 QSRank approximation

In order to compute fast retrieval, the QsRank will be approximated by a lighter formula.
First, only the top k dimensions of the PCA projection will be used (k will be defined afterward in the next section). Thus, $NN(q, \epsilon)$ becomes $NN(q^k, \epsilon)$ and $S(h)$ becomes $S(h^k)$ where x^k is the k-top dimensions of a vector $x \in \mathbb{R}^d$. This approximation seems legit since the aim of the PCA is to find the dimensions where most of the information is kept.

$$\text{QsRank}(q, h, \epsilon) \approx \frac{\int_{NN(q^k,\epsilon)\cap S(h^k)} p(y^k)dy^k}{\int_{NN(q^k,\epsilon)} p(y^k)dy^k}$$

Another approximation is to replace the ϵ-ball by an ϵ-hypercube:

$$NN(q^k, \epsilon) \leftrightarrow HC(q^k, \epsilon) = \{y^k \in \mathbb{R}^k \text{ s.t. } \forall i \in [1; k], |y_i^k - q_i^k| < \epsilon\}$$

Moreover, since the PCA produces uncorrelated projections, each dimension of $p(y)$ is supposed to be independent. The QsRank approximation then becomes:

$$\text{QsRank}(q, h, \epsilon) \approx \prod_{i=1}^{k} \frac{\int_{|y_i^k - q_i^k| < \epsilon, y_i^k h_i^k > 0} p(y_i^k)dy_i^k}{\int_{|y_i^k - q_i^k| < \epsilon} p(y_i^k)dy_i^k} = \prod_{i=1}^{k} \mathbb{P}(y_i^k \in S(h_i^k) | y_i^k \in HC(q_i^k, \epsilon))$$

The last approximation is to consider that the y are generated from a uniform law. This assumption accelerates a lot the computation and seems to work quite well in accordance with the authors of [1]. The final formula thus becomes:

$$\begin{aligned} \text{QsRank}(q, h, \epsilon) &\approx \prod_{i=1}^{k} \frac{\int_{|y_i^k - q_i^k| < \epsilon, y_i^k h_i^k > 0} dy_i^k}{\int_{|y_i^k - q_i^k| < \epsilon} dy_i^k} \\ &\approx \prod_{i=1}^{k} \text{clamp}\left(\frac{1}{2}\left(1 + \frac{h_i^k q_i^k}{\epsilon}\right), [0; 1]\right) \end{aligned}$$

Opposed to the Hamming distance, this measure has many substantial advantages :

- The radius ϵ is took into account
- Since it is a product, if one of the component of the hash code induces a null probability, the whole QsRank becomes null. Some sets are thus not explored whereas Hamming distance methods would have.

3 Implementation

3.1 Efficient computation

Once the query q is given, the logarithmic QsRank is computed for the 2^k different hash codes. This results in $\mathcal{O}(k)$ logarithm computation and $\mathcal{O}(k2^k)$ additions. Moreover, only hash codes with non zeros probabilities have to be computed, which accounts for only 15% of the data points according to the experiments made in [1].
The idea of the retrieval is also to make a pre-selection with QsRank computed only on the first K_1 bits. Then, the L first buckets are selected and more accurate re-ranking (on $K_1 + K_2$ bits) is computed to select the ϵ-neighbors.

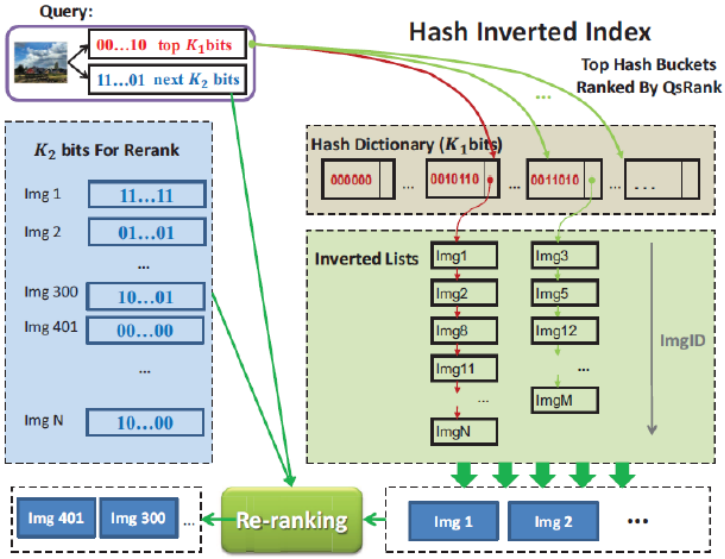

Figure 1: Retrieval procedure: Buckets ranking (K_1 bits). Neighbors re-ranking ($K_1 + K_2$ bits). Image credits from [1]

3.2 Retrieval procedure

Input : input data $X \in \mathbb{R}^{n \times d}$, query q, radius ϵ
Parameters: hash codes length K_1 and K_2, number of buckets L.
Precomputation
$Y = PCA(X)$
for $i = 1$ .. $K_1 + K_2$ **do**
  Compute the logarithmic sub-QsRanks : $\log\left(\text{clamp}\left(\frac{1}{2}\left(1 + \frac{h_i^k q_i^k}{\epsilon}\right), [0; 1]\right)\right)$
end
ϵ-search
Sort the buckets by K_1-QsRank.
Pick the L^{th} first buckets.
Sort the buckets hash codes by $(K_1 + K_2)$-QsRank.
Return the ϵ-neighbors.

3.3 Implementation details

Our implementation use *Matlab* with the statistic toolbox. The script `main.m` contains an exemple search in data points loaded from `data/feat_oxc1_hesaff_sift.bin`[1].

The script `recall_graph.m` was used to generate the two recall-loss graph that are presented in this report.

And the script `demo2d.m` show the first two steps of the algorithm (PCA and K_1-QsRank separation) in $2D$

4 Results

We have implemented the algorithm the way it is explained above. The parameters used are the one proposed in [1]. They are summed up in the table below.

K_1	K_2	L
16	48	50

Table 1: Default parameters used for the algorithm.

Then, in order to evaluate its efficiency, we have generated many queries and compared the results obtained by the algorithm and the true neighbors with the *recall loss*:

$$r_{loss} = \frac{|\text{found neighbours}|}{|\text{true neighbours}|}$$

The results can be observed on figure 2

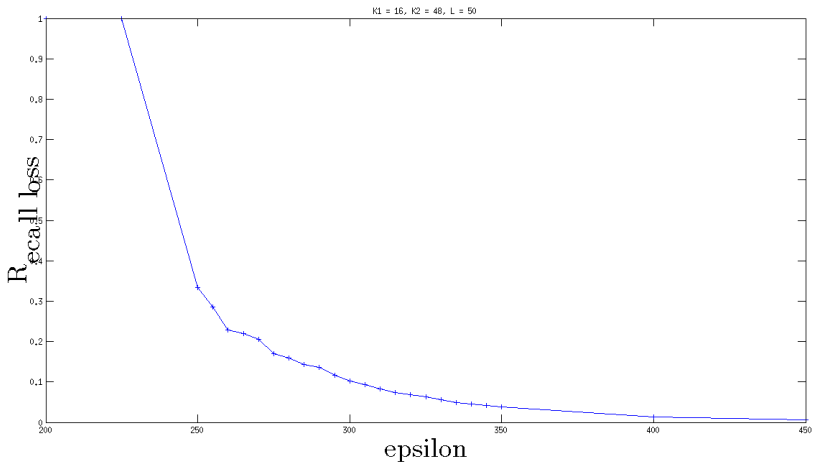

Figure 2: Recall Loss function of epsilon.

This graph shows that ϵ needs not to be too high in order to keep the approximations made in the previous sections correct. Otherwise, the results are quickly inaccurate.
We can also observe, in figure 3, the loss as a function of K_1: when K_1 get higher, the number of buckets increase and the average number of points per bucket decrease. With a fixed number L of buckets selected, the number of points that are considered irrelevant by the algorithm after the K_1-QsRanking increase

[1]Download from http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/feat_oxc1_hesaff_sift.bin.tgz and unzip to data/

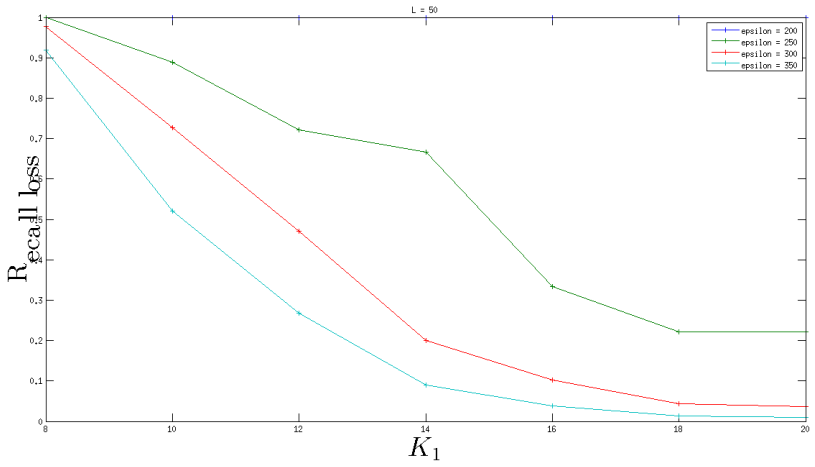

Figure 3: Recall Loss function of K_1.

when K_1 decrease. This explain why the loss is inversely proportional to K_1.

Although, the algorithm seems to have good speed performances ($\sim 0.2s$ according to [1]), ours is not as fast. This could be explained by the use of the *Matlab* software and the hashmap generations that seems to take too much time (*Matlab* require our algorithm to convert back and forth between table of -1 and 1 and *uint64* to represent the hashcodes. This shouldn't be necessary).

5 Conclusion

With this article, we have discovered the field of large-scale-ϵ-neighbors search. A field that becomes more and more significant because of the huge increase of available data on the internet. The method presented in [1] differs a lot from the other state-of-the-art ones by not using the Hamming distance for the retrieval. It induces a better precision for the search. Moreover, the number of bits used is much smaller (only 64 bits) which reduces memory cost speed up the process. However, this number is not very movable since it depends on the PCA and thus on the initial dimension of the space.
Finally, this method is powerful by its simplicity and efficiency. Many approximations are made to make this algorithm fast but it is still very efficient as it is equal to other state-of-the-art methods.

References

[1] L. Zhang X. Zhang and H-Y. Shum. Qsrank: Query-sensitive hash code ranking for efficient epsilon-neighbor search. *CVPR*, 2012.