Course Project: #4: Index Quantization

Group Member

Haozhong Zheng, hz2675@nyu.edu

Description

In course final project, I'm going to work on precomputed index score quantization using my own system from HW2 and HW3.

There are multiple approaches for floating-point quantization or compression, including both lossless compression and lossy compression:

1. Primitive approaches. Such as computing value range of index scores, then mapping them into 1 byte; or converting floating-points to integers with some decimals, then using index compression.
2. General compressor. Using third-party library like HDF5 to store index scores data, then using some general compressor like gzip.
3. I may do search for papers about state-of-the-art lossy compression algorithms and try to implement some of these.

Things to achieve

In this project, I would use third-party library or implement algorithms by myself. Therefore, my own system from HW2 and HW3 have to be drastically refactored in order to be more generic.

I'll try to find recent papers about floating-point compression and compare various different approaches about their speed, compression ratio and query quality.

Feedback:
I think you should focus more on 1. There are a variety of approaches, and they are not all primitive. You can do linear approximation, or logarithmic, or a variety of other things. And you can vary the resulting size in bits of the quantized values, say from 4 to 12 bits, and see what happens.