Proposal for final project: explore use splay tree to improve clustering

Chao Ren

My plan to the final project is an experimental research. It is exploring use splay tree to improve k means clustering.

The path I get to this idea is: first, I went to my advisor look for some idea about how data structure can work with machine learning, then she showed me the really large image retrieval problems, which means finding the closest feature vector of an image in millions of feature vectors. To solve this, if we have a clustering to these vectors, it may help to accelerate finding the vector, because the clustering helps gathering similar vectors together. For the clustering , I find a paper using RB tree and min heap to accelerate k means clustering, which is a classic clustering algorithm. So what comes to my mind is that since RB tree can accelerate clustering, then how will splay tree perform on the clustering algorithm, given that they are both self-balance tree? Then I searched internet and found few information about this. So I think this might be a possible idea to explore, and even I may not get the clustering accelerated, it worth explore that why it cannot.

For the already known information, k means algorithm is a partition algorithm. By this algorithm, we can divide the vector set into k groups by iterations. These groups should try to make vectors in a group as close as possible, and vectors in different group as far as possible. This algorithm takes two steps: first, choose k centroids of initial clusters, then add vectors to the cluster with closest centroid. Second, recalculate the centroids, recalculate the distance and still add vectors into closest cluster. Iterate by this way until no more vector changes its cluster. This is time consuming because I need to calculate distance between every vector to every cluster. But in fact, some distances doesn't need recalculate if no vector change in these clusters. So in this way if we optimize it by remove unnecessary recalculation, then it should be faster. So from Kumar, 2011, we know RB tree can accelerate k means clustering at least by the mean of time(they acknowledge that this acceleration brings more space consuming). And since we learned splay tree in class, I am curious how will splay tree do to this, will splay tree do better or same, or even make it worse? And why?

So that is what I want to investigate: if I replace RB tree with splay tree in this k means clustering acceleration algorithm, will anything be different or even is it possible to implement?

Reference

Rajeev, K., Rajeshwar, P.,& Joydip, D.,(2011) Enhanced K-Means Clustering Algorithm Using Red Black tree and Min-Heap.

International Journal of Innovation, Management and Technology, 2(1), 49-54. Matthijs, D.,(2018) *Indexing 1M vectors*, Retrieved from

https://github.com/facebookresearch/faiss/wiki/Indexing-1M-vectors