Xinyu Jiang

02/26/2021

109036441

Paper Review for Mar 2

Paper Review

The paper "Multi-Probe LSH: Efficient Indexing for High-Dimensional Similarity Search" introduces a new technique to solve the overuse of time and space in content-related search systems: multi-probe locality sensitive hashing (LSH). The rest of the paper will analyze this method with its strengths and weaknesses.

Traditional LSH methods usually need a large amount of hash tables to protect the searching accuracy. However, it decreases the searching efficiency and abuses the space. Traditional LSH use hash functions to map similar data to the same hash bucket with a high probability. After data has been put in the bucket, LSH method will sort the data based on its distance and return the top-N data (Lv, 2007). However, with this method, it needs a large amount of hash tables to ensure that most of the neighboring data is covered, which requires a lot of space. Multi-probe LSH is the technique that solves the problem. It can intelligently detect multiple buckets in the hash table that may contain query results, so that its sufficiently improved space and time efficiency compare with traditional LSH (Lv, 2007). Based on the nature of traditional LSH, if the similar data is not mapped to the same bucket, then it will go into its surrounding bucket. Then, for multi-probe LSH, the goal is to locate these neighboring buckets in order to increase the chance of finding neighboring data, so that decreasing the searching time.

Even though multi-probe LSH has a lot of advantages, there are still some limitations that constraint its usage. The first one is about data type. According to the core of multi-probe LSH, it

uses a carefully derived probing sequence to obtain multiple hash buckets that are similar to the query data (Lv, 2007). However, if the relationship between each data is not strong or even no relation, it will be hard for the algorithm to find the correlation between each other, so that multiple-probe LSH has no difference with traditional LSH algorithm. Another limitation is also about data type, LSH hopes to use hash tables to normalize similar elements into the same bucket. After multiple probes, it is equivalent to multiple hashes, which increases the number of calculations and increases time consumption

To conclude, Multi-probe Locality Sensitive Hashing does effectively reduce the time cost in searching high-dimension data, but there is also limitation for that method, especially in terms of data selection. Therefore, to correctly optimize the algorithm, programmers need to carefully select the data.

References

1. Lv, Q., Josephson, W.K., Wang, Z., Charikar, M., & Li, K. (2007). Multi-Probe LSH: Efficient Indexing for High-Dimensional Similarity Search. VLDB.