A Survey of LSH for Similarity Search

Zeyan LI 2018310816 **ChenCheng Xu 2018310851**

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

2 Methods of LSH

2.1 Multi-Probe LSH

Traditional LSH schemes directly return the objects in the exactly collision buckets. Multi-probe LSH [1] try to probe more buckets in one hash table to reduce the number of hash tables while maintains similar performance. Formally speaking, multi-probe LSH probes a sequence of buckets which are from the collision bucket and a sequence of perturbations, as Table 1 shows.

Table 1: Difference between basic LSH and multi-probe LSH

Scheme	Query
Basic	$g(q) = (h_1(q), h_2(q), ..., h_M(q))$
Multi-Probe	$g(q)+\Delta^{(i)}, i=1, 2, ..., T$, $\Delta^{(i)}=(\delta_1^{(i)}, \delta_2^{(i)}, ..., \delta_M^{(i)})$

Then we are going to introduce how the perturbation sequences are constructed.

2.1.1 Step-Wise Probing Sequence

Given the properties of LSH, similar objects are more likely in the close buckets. This motivates the step-wise probing sequence, which firstly probes all 1-step perturbations, then all 2-step perturbations, and so on. There are $L\times\binom{M}{n}\times 2^n$ n-step perturbations in total, where L denotes the number of hash tables and M denotes the number of compound hash functions in each table.

2.1.2 Query-Based Probing Sequence

Step-wise probing just consider all coordinates to be equally likely. However, according the properties LSH, some perturbations are more likely than others.

We consider this hash function: $h(q) = \lfloor \frac{a\cdot q+b}{w} \rfloor$, where w is a fixed hyper-parameter, a is drawn from standard Gaussian, and b is uniformly drawn from $[0, w)$. For two objects

Preprint. Work in progress.

References

[1] Q. Lv, W. Josephson, Z. Wang, M. Charikar, and K. Li, "Multi-probe lsh: efficient indexing for high-dimensional similarity search," in *Proceedings of the 33rd international conference on Very large data bases*. VLDB Endowment, 2007, pp. 950–961.