Efficient Distributed Locality Sensitive Hashing

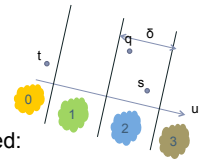
Bahman Bahmani[1], Ashish Goel[1], Rajendra Shinde[1]

Overview

- Similarity search (SS)
  - data mining, content based search, etc.
- **Layered LSH**: network efficient implementation on (Key, Value) based distributed frameworks
  - MapReduce
  - Active Distributed Hash Tables
    - Storm, S4
- Exponential improvement in network cost

Locality Sensitive Hashing

- SS with distance thresholds l, c*l
  - Similar: distance at most l
  - dissimilar: at least c*l
- Construct hash functions
  - nearby points collide:
    - E.g. q and s
  - distant points separated:
    - E.g s and t
- Accuracy: need several hash tables
  - Each table uses concatenated hashes from many LSH functions as keys

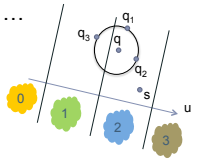


Basic LSH

- Insert data points in L hash tables, L = $O(n^{1/c})$
- Query hashed using same functions
- Search buckets to which query is hashed to
- Simple MR implementation (**Simple LSH**):
  - Emit (H, x) if x falls into bucket H
  - Reducer implements SS in its bucket
- Network cost is O(L) per query and data point

Entropy LSH

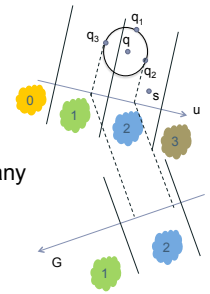
- **Single** hash table
- Choose random offsets q_1, q_2, q_3, … near each q
- Search buckets offsets map to
- Network cost: #offsets per query
- $O(n^{2/c})$ offsets for accuracy



Layered LSH

- Distribute buckets so that nearby points are likely to be on the same machine
- Rehash buckets via an additional layer of LSH (G)
  - For data point x, emit (GoH(x)), (x,H(x)))
  - For queries q, emit (GoH(q_i),q) for each offset q_i
  - Reducer generates offsets for each query and searches their buckets

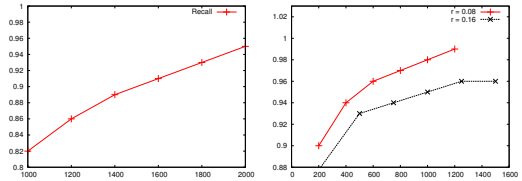
- Query offsets are near
  - Buckets differ in few bits
  - Sent to few machines
  - **Simple LSH** may send every offset to a different machine
- Distant points: buckets differ in many bits.
  - Unlikely to be sent to the same machine

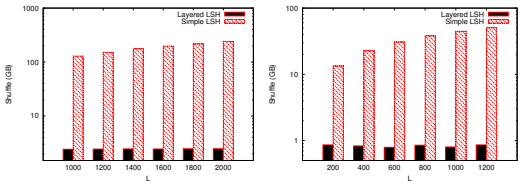


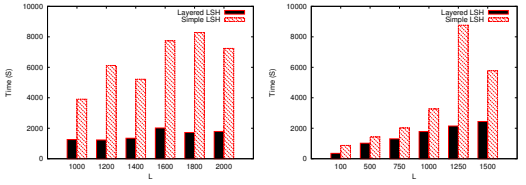
- **Results:**
  - each query maps to at most sqrt(log n) machines, an **exponential** improvement
  - Points at a constant separation are sent to distinct machines

Experiments

- Comparison with **Simple LSH**
- Similarity Join experiments on 16 node Hadoop cluster
- Data:
  - Random, d=100
    - 1M data, 100K query
    - Queries artificially planted near data points
  - Tiny Image dataset
    - 384 GIST descriptors per image
    - 1M data, 200K queries







1: Stanford University.