Two-Stage Hashing for Fast Document Retrieval

Hao Li* Wei Li◆ Heng Ji*

****Computer Science Department, Rensselaer Polytechnic Institute***

◆IBM T.J. Watson Research Center

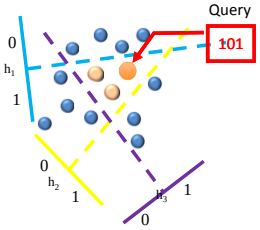
(lih13, jih)@rpi.edu; weiliu@us.ibm.com

Document Retrieval in Big Data

- Traditional IR Methods
  - Memory consuming: represent documents in a vector space
  - Time consuming: cosine similarity calculation
  - Infeasible for large-scale datasets
- Hashing Methods
  - Compact: represent documents as binary codes (e.g., d_1 -> "10100100")
  - Efficient: hamming distance calculation, hash table lookup
  - Scalable to massive datasets

1

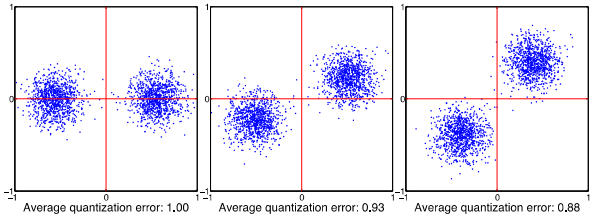
Locality Sensitive Hashing



- Intuition
  - If two data points are close, then after the "projection + binarization" operation they will remain close.
- Advantages:
  - Randomized Hashing: time efficient for search
  - Very high hash table lookup success rate (100% with more than 2 tables)
- Drawback:
  - Inadequate search precision

2 (Andoni and Indyk, 2008) (Liu, 2012)

Iterative Quantization

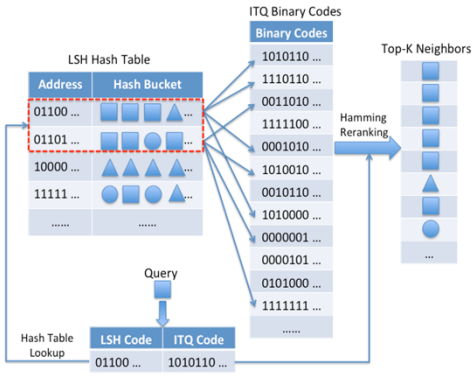


(a) PCA aligned. (b) Random Rotation. (c) Optimized Rotation.

- Intuition: learn the binary codes achieving the lowest quantization error by seeking a proper rotation of zero centered projected data
- Advantages:
  - Well approximate real-valued data with binary codes
  - High search precision via Hamming distance ranking
- Drawback:
  - poor hash lookup success rate with longer bits (18.47% with 384 bits)

3 (Gong et al., 2013)

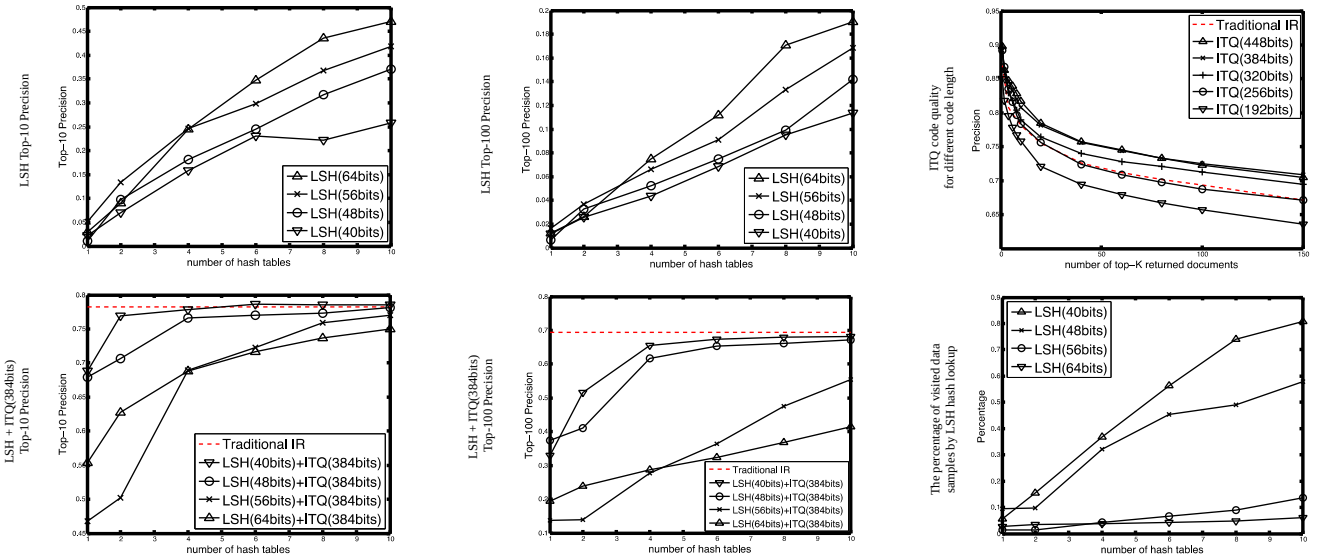
Two-Stage Hashing Framework



- LSH for neighbor candidate pruning; ITQ for effective re-ranking
- LSH captures term similarity; ITQ captures topic similarity
- Advantages:
  - High hash lookup success rate is attained by the LSH stage
  - High search precision due to the ITQ re-ranking stage
  - Scan only a small portion of an entire dataset
  - Integrate two similarity measures

4

Experiment Results



- Comparable search accuracy with the traditional IR method
- An order of magnitude speedup in search time (1/30 of traditional IR search time)

5