CS522 Advanced Database Systems

Hash Tree

Chengyu Sun
California State University, Los Angeles

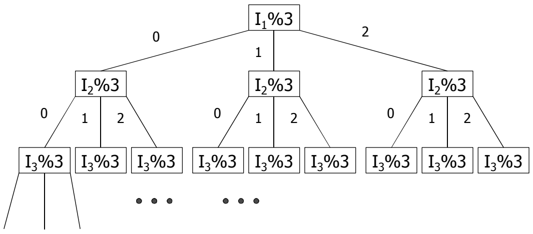
Hash Tree

- There is a hash function associated with each internal node
- Which branch to follow is determined by the hash value

A Hash Tree for 3-Itemsets …

- A 3-itemset can be written as $\{I_1,I_2,I_3\}$, where I_1 is the first item, I_2 is the second item, I_3 is the third item

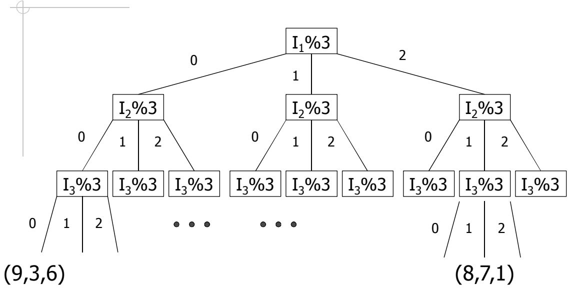
… A Hash Tree for 3-Itemsets



Insertion (i.e. Hashing) …

- Suppose we want to insert (i.e. hash) the following 3-itemsets into the tree
  - (9,3,6)
  - (8,7,1)

… Insertion (i.e. Hashing) …



… Insertion (i.e. Hashing)

- (9,3,6) is inserted into the left-most leaf because
  - At level 1, 9%3=0
  - At level 2, 3%3=0
  - At level 3, 6%3=0
- Similarly, (8,7,1) is inserted (i.e. hashed) to the leaf following the path 2-1-1

Support Counting Using a Hash Tree …

- Suppose we want to do support counting for C_k (i.e. candidate k-itemsets)

… Support Counting Using a Hash Tree …

- Create a hash tree and hash all the candidate k-itemsets to the leaf nodes of the tree
- For each transaction, generate all k-item subsets of the transaction
  - E.g. for a transaction {1,2,3,4}, the 3-item subsets are {1,2,3}, {1,2,4}, {1,3,4}, and {2,3,4}

… Support Counting Using a HashTree

- For each k-item subset, hash it to a leaf node of the hash tree, and check it against the candidate k-itemsets hashed to the same leaf node. If the k-item subset matches a candidate k-itemset, increment the support count of the candidate k-itemset

Advantage of Support Counting Using Hash Tree

- Each k-item subset is only checked against the candidates hashed to the same leaf instead of all candidates

Disadvantage of Support Counting Using Hash Tree

- Creating the hash tree takes some coding