UNIVERSITY of DELAWARE

Implementation Issues

(CPEG 457/657: Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

UNIVERSITY of DELAWARE

Query Processing Modes

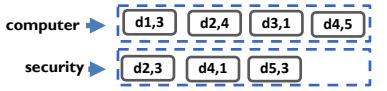
- Fetch the inverted list for all query terms
- Perform set operations to get the subset of docs that satisfy the Boolean condition
  - Conjunctive mode (AND)
  - Disjunctive mode (OR)
  - E.g., Q="computer security"
    - computer: d1, d2, d3, d4
    - security: d2, d4, d6
    - Results:
      - Conjunctive mode: compute relevance scores for {d2,d4}
      - Disjunctive mode: compute relevance scores for {d1,d2,d3,d4,d6}

UNIVERSITY of DELAWARE

Query Processing

- Given a query and indexes, we need to
  - Traverse the indexes
  - Compute relevance scores of documents
  - Rank documents

Query: computer security

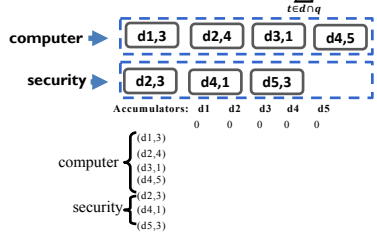


How to traverse the indexes?

UNIVERSITY of DELAWARE

Term-at-a-time (TAAT)

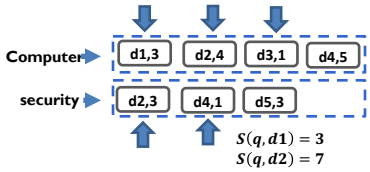
Query: computer security $S(q,d) = \sum_{t \in d \cap q} c(t,d)$



UNIVERSITY of DELAWARE

Document-at-a-time (DAAT)

Query: computer security $S(q,d) = \sum_{t \in d \cap q} c(t,d)$



$S(q,d1) = 3$
$S(q,d2) = 7$
$S(q,d3) = 1$

UNIVERSITY of DELAWARE

Further Improving Efficiency

- Keep only the most promising accumulators
- Sort the inverted list in decreasing order of weights and fetch only N entries with the highest weights
- Pre-compute as much as possible
- Dynamic pruning methods