CMPS 561
Boolean Retrieval

Ryan Benton
Sept. 7, 2011
Algorithms for Intersection
Algorithms – Basic Intersection (aka Merging)

- Intersect(p1, p2)
  - answer \(\leftarrow\) {}
  - While \((p1 \neq \text{NIL})\) and \((p2 \neq \text{NIL})\) Do
    - if \(\text{docID}(p1) = \text{docID}(p2)\)
      - Then ADD(\(\text{answer}\), \(\text{docID}(p1)\))
        - p1 \(\leftarrow\) next(p1)
        - p2 \(\leftarrow\) next(p2)
    - Else if \((\text{docID}(p1) < \text{docID}(p2))\)
      - Then p1 \(\leftarrow\) next(p1)
      - Else p2 \(\leftarrow\) next(p2)
  - Return answer
Algorithms – Intersection

• **Complexity:** $O(x + y)$
  – For any given two posting lists
    • List A has size $x$
    • List B has size $y$
  – Note, this is upper bound.

• **Formally, Complexity:** $\Theta(N)$
  – $N$ can be either
    • Number of documents in collection
  – Note, this is a tight bound.
Observation

• In many cases, Boolean queries
  – Conjunctive in nature
• Allows for a possible improvement based on posting size (term frequency)
Algorithms – Conjunctive Query Merging

- IntersectConjunct(\(t_1, t_2, ..., t_z\))
  - Terms \(\leftarrow\) SortByIncreasingFrequency((\(t_1, t_2, ..., t_z\)))
  - Results \(\leftarrow\) postings(first(Terms))
  - Terms \(\leftarrow\) rest(Terms)
  - while (Terms != NIL) and (Results != NIL) Do
    - Results \(\leftarrow\) Intersect(result, postings(first(Terms)))
    - Terms \(\leftarrow\) rest(Terms)
  - Return Results
Why?

• By using least frequent term
  – All results guaranteed to be no larger than least frequent term

• In practice
  – The ‘intermediate’ list always places upper bounds on the size.
References

• Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *Introduction to Information Retrieval*, Chapter 1, 2008.


• Vijay V. Raghavan’s Notes/Lecture Material
  – Material in Slides ued with permission