Midterm summary

March 12, 2008

All material up to but not including today will be covered on exam out Wednesday, March 26 and due Friday, March 28, 2008

1

Major topics (highlights) to date

- Models
  - Modeling queries and documents
    - Satisfying, scoring and ranking w.r.t query
  - Vector models
    - LSI
  - Link analysis for ranking
- Evaluating retrieval systems
- Algorithms and data structures
  - Index structures
  - Evaluating queries using indexes
  - Building indexes

2

Reading assigned to date in *Introduction to Information Retrieval*

- Chapter 1 – all sections
- Chapter 2 – all sections
- Chapter 3 – section 3.1: structures for dictionaries
  - Rest on wildcard queries and spelling correction
- Chapter 4 – all sections
- Chapter 5 – section 5.1: Heaps' and Zipf's laws
  - Rest of sections are for today, not on test
- Chapter 6 - all sections except:
  - 6.1 zone scoring
  - 6.4.4 advanced dealing with document length in scoring

3

Continuation of Reading assigned to date in *Introduction to Information Retrieval*

- Chapter 8 – all sections
- Chapter 18 – all sections
- Chapter 19 –all sections except:
  - 19.5 estimating relative sizes of search engine indexes
  - 19.6 finding near-duplicate pages
- Chapter 21 - all sections

4

Chapter 7

- Assigning now as included chapter for test
  - Except 7.1.6: uses clustering
- Mostly summarizes issues we have touched on
  - Primarily use as review
  - Just understand basics for techniques we haven't discussed

5

Major ideas Chapter 7

- Approximating k best documents for a query
  - Ordering posting lists by a score of doc and limiting number of documents consider from each list
    - Global score – e.g. PageRank
    - Score of doc for term
      - Won't have same doc. order all posting lists
      - Several techniques suggested – understand basics
- Using proximity of terms in scoring

6