Remarks on index building and query processing

1

Remarks: Index Building

- Google's pools of machines not specialized - can be assigned to do any one of set of tasks - e.g. forward index, invert index - whenever become free
- Aggregate Information on terms, e.g. document frequency, also needs to be computed as compute index
  - store w/ dictionary
- May not actually keep every occurrence, maybe just first k.
  - Early Google did this for k=4095. Why?
- What happens if dictionary not fit in main memory as build inverted index?

2

Query Processing
High-level parallelization by Google

circa 2002

- Enter query -> DNS-based directed to one of geographically distributed clusters
  - Load balance and round-trip time
- w/in cluster, query directed to 1 Google Web Server (GWS)
  - Load balance
- Query distributed to pools of machines, each pool handling subset of docs
  - Load sharing
- Query directed to 1 machine w/in each pool
  - Load balance

3

Web query processing: limiting size

- For Web-scale collections, may not process complete posting list for each term in query
  - at least not initially
- Need docs sorted first on global quality, e.g. PgRank,
- only take first k for some k that depends on query and how many want to be able to return
  - Google:
    - 1000 max returns;
    - 32 words max query size
  - How query affect k?
  - Flaws w/ partial retrieval from each list?

4

More observations

- Consider Google behavior
  - Stop words
  - Stemming

5

Compression

on board

6