Massive Text Indices

Johannes Fischer (TU Dortmund)
Peter Sanders (KIT)

Johannes Fischer

tu technische universität dortmund
KIT Karlsruhe Institute of Technology
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Large Data Sets

Large Data Sets

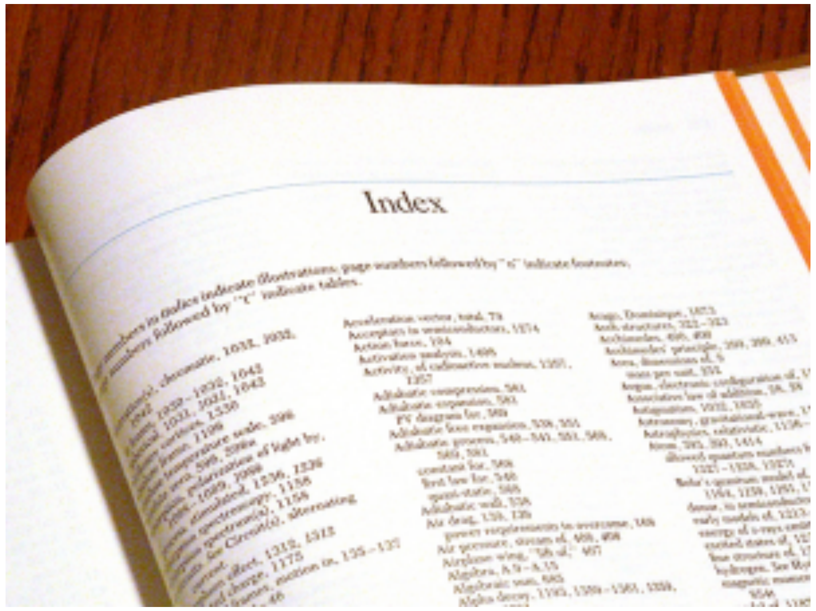
Large Data Sets

Large Data Sets

- in particular: **texts**
- **indexing** required (for fast search)

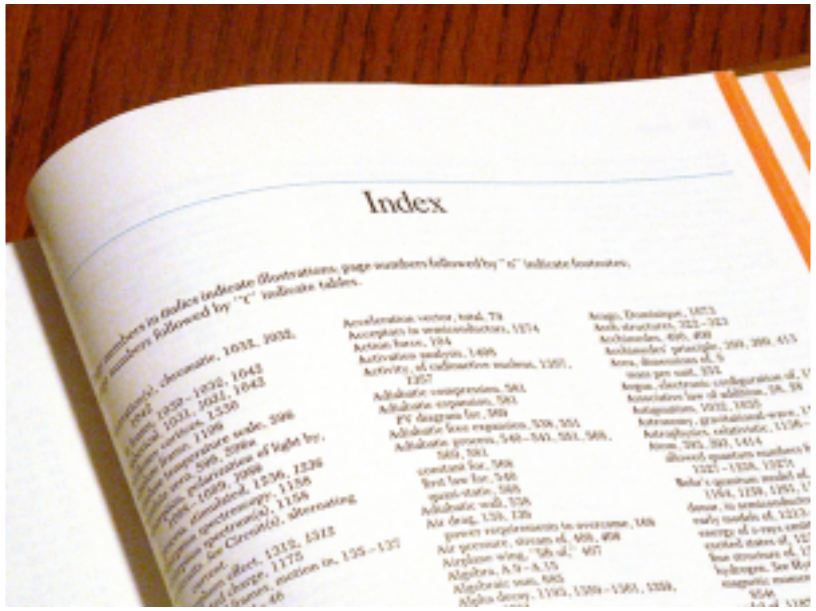
Text Indexing

- Problem: **preprocess** a text *T* for **fast** pattern queries (without scanning *T* again)



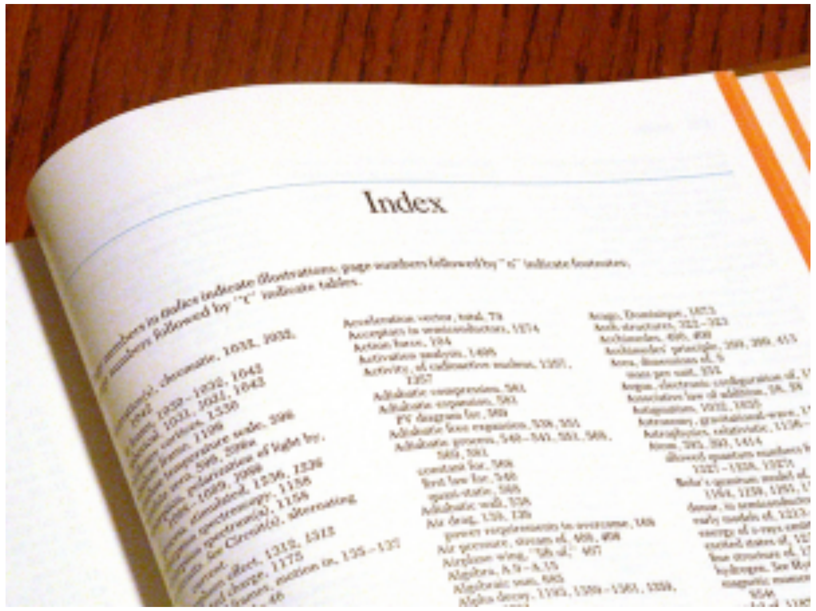
Text Indexing

- Problem: **preprocess** a text *T* for **fast** pattern queries (without scanning *T* again)
  - just preprocess once (*T* static)



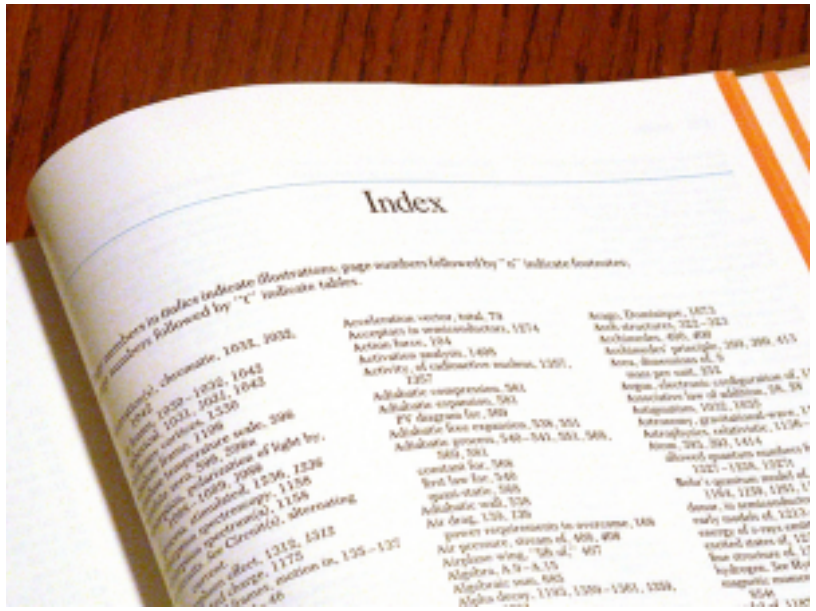
Text Indexing

- Problem: **preprocess** a text *T* for **fast** pattern queries (without scanning *T* again)
  - just preprocess once (*T* static)
  - many pattern searches



Text Indexing

- Problem: **preprocess** a text *T* for **fast** pattern queries (without scanning *T* again)
  - just preprocess once (*T* static)
  - many pattern searches
  - e.g. PDFs



3 Issues:

3 Issues:

Construction

3 Issues:

Construction

Storage

3 Issues:

Construction

Storage

Querying

Project Goals

- Computation/Storage:
  - distribution over hundreds of nodes
  - nodes make full use of computing power
    - external memory
    - succinct data structures
    - shared memory parallelism (CPU/GPGPU)
- Query Distribution?

Project Goals

- Computation/Storage:
  - distribution over hundreds of nodes
  - nodes make full use of computing power
    - external memory
    - succinct data structures
    - shared memory parallelism (CPU/GPGPU)

combination?

- Query Distribution?