Index Creation

Sven Fiergolla

23. März 2018

Übersicht

Einführung

Effiziente Suche über:

Sammlung von Büchern

das Web

Große Datenmengen

zu viel für Main Memory!

Einführung

Effiziente Suche über:

Sammlung von Büchern

das Web

Große Datenmengen

zu viel für Main Memory!

Einführung

Effiziente Suche über:

Sammlung von Büchern

das Web

Große Datenmengen

zu viel für Main Memory!

Einführung

Typische Systemeigenschaften (stand 2018)

- *clock rate* 2-4 GHz, 4-8 Kerne
- *main memory* 4-32 GHz
- *disk space* $\leq$ 1 TB SSD oder $\geq$ 1 TB HDD
  - HDD (hard disk drive)
    - *average seek time* zwischen 2 und 20 ms
    - *transfer time* 150 - 300 MB/s
  - SSD (solid state disk)
    - *average seek time* zwischen 0.08 und 0.16 ms
    - *transfer time* Lesen: 545 MB/s, Schreiben: 525 MB/s

Einführung

Typische Systemeigenschaften (stand 2018)

- *clock rate* 2-4 GHz, 4-8 Kerne
- *main memory* 4-32 GHz
- *disk space* $\leq$ 1 TB SSD oder $\geq$ 1 TB HDD
  - HDD (hard disk drive)
    - *average seek time* zwischen 2 und 20 ms
    - *transfer time* 150 - 300 MB/s
  - SSD (solid state disk)
    - *average seek time* zwischen 0.08 und 0.16 ms
    - *transfer time* Lesen: 545 MB/s, Schreiben: 525 MB/s

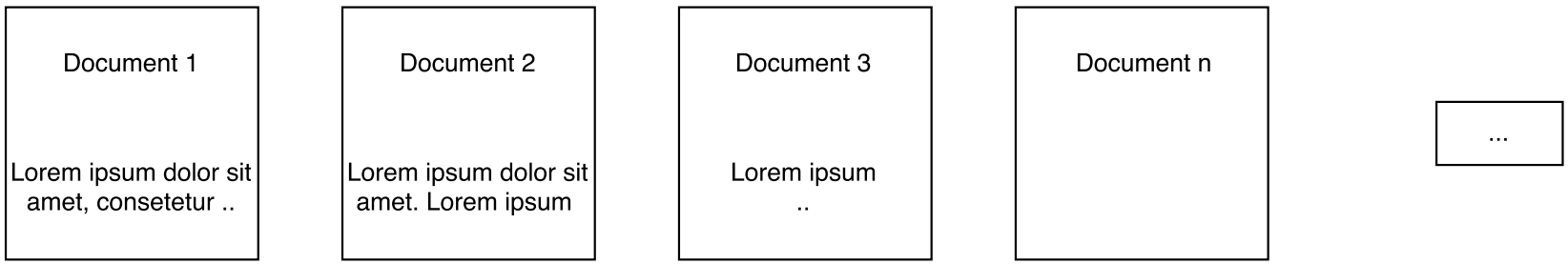
Einführung

Typische Systemeigenschaften (stand 2018)

- *clock rate* 2-4 GHz, 4-8 Kerne
- *main memory* 4-32 GHz
- *disk space* $\leq$ 1 TB SSD oder $\geq$ 1 TB HDD
  - HDD (hard disk drive)
    - *average seek time* zwischen 2 und 20 ms
    - *transfer time* 150 - 300 MB/s
  - SSD (solid state disk)
    - *average seek time* zwischen 0.08 und 0.16 ms
    - *transfer time* Lesen: 545 MB/s, Schreiben: 525 MB/s

hardware constraints

Indizierung einer Sammlung von Daten auf der Festplatte

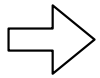


Zugriffszeit auf Festplatte als Bottelneck

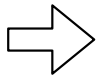
Index Creation

geeignete Datenstruktur um Zugriff auf die Festplatte zu minimieren

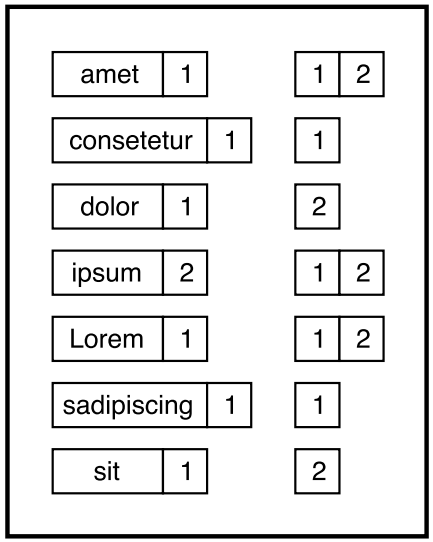
TermID / DocID	
Lorem	1
ipsum	1
consetetur	1
sadipscing	1
amet	1
ipsum	1
Lorem	2
ipsum	2
dolor	2
sit	2
amet	2
...	2



TermID / DocID	
amet	1
amet	2
consetetur	1
dolor	2
ipsum	1
ipsum	2
Lorem	1
Lorem	2
sadipscing	1
sit	2
...	...



TermID , doc frequenz / Postings List



Index Creation - hardware constraints

in der Regel übersteigt die Datenmenge der Dokumente den Main Memory.

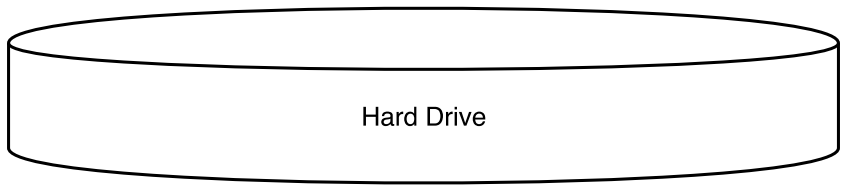
Reuters-RCV1 benötigt ca 0.8 GB für die termID/DocID Paare, bereits die DBLP übersteigt die Dokumentenanzahl, besitzt jedoch weniger Terme.

Blocked sort-based indexing (BSI)

todo: erklären usw

Blocked sort-based indexing (BSI)

Zusammenführen von Postings Lists bei block sort-based Indexing

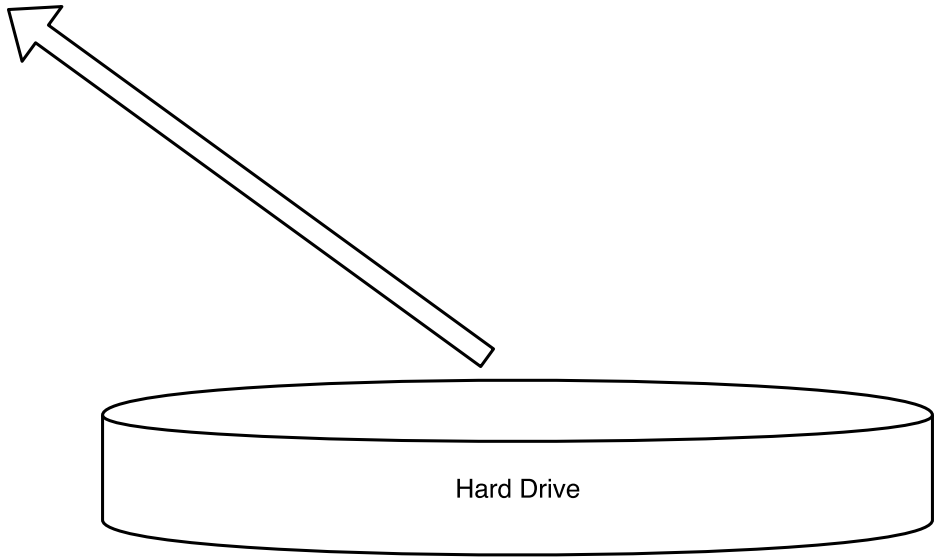


Blocked sort-based indexing (BSI)

Zusammenführen von Postings Lists bei block sort-based Indexing

amet	d1,d2
dolor	d2
Lorem	d1
ipsum	d2

consetetur	d3, d4
Lorem	d3
ipsum	d3
sit	d4

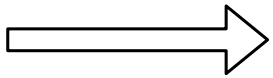


Blocked sort-based indexing (BSI)

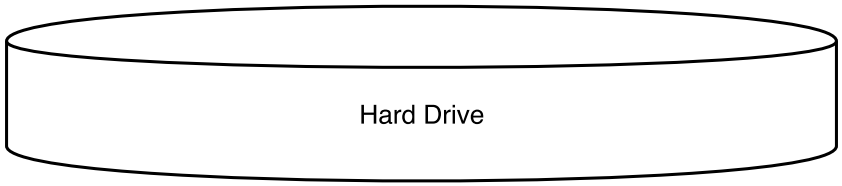
Zusammenführen von Postings Lists bei block sort-based Indexing

amet	d1,d2
dolor	d2
Lorem	d1
ipsum	d2

consetetur	d3, d4
Lorem	d3
ipsum	d3
sit	d4



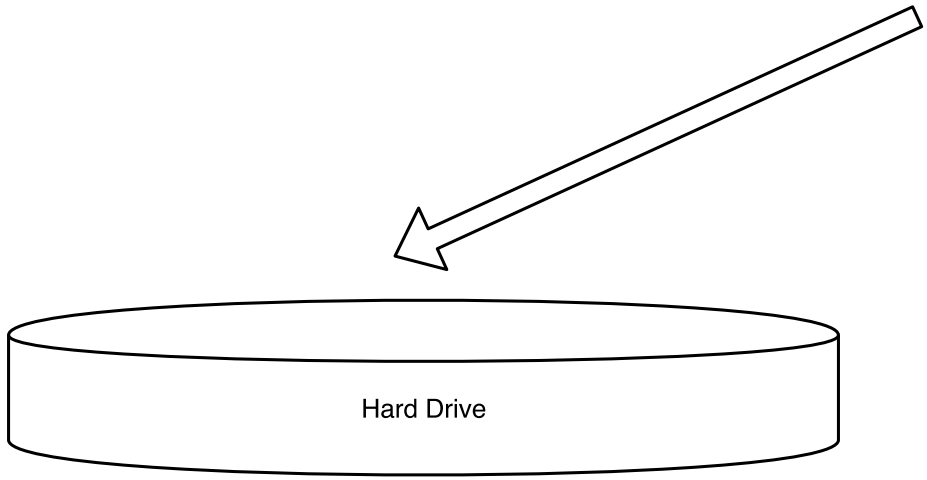
amet	d1,d2
consetetur	d3,d4
dolor	d2
Lorem	d1, d3
ipsum	d2, d3
sit	d4



Blocked sort-based indexing (BSI)

Zusammenführen von Postings Lists bei block sort-based Indexing

amet	d1,d2
consetetur	d3,d4
dolor	d2
Lorem	d1, d3
ipsum	d2, d3
sit	d4



Single-pass in-memory indexing (SPIMI)

Quellen

- Shannon, C. E. „A Universal Turing Machine with Two Internal States.“ Automata Studies. Princeton, NJ: Princeton University Press, pp. 157-165, 1956. [1]
- Wolfram Research and Wolfram, S. „The Wolfram 2,3 Turing Machine Research Prize.“ [2]
- Wolfram, S. A New Kind of Science. Champaign, IL: Wolfram Media, pp. 706-711 and 1119, 2002.

[1] `http://www.sns.ias.edu/~tlusty/courses/InfoInBio/Papers/Shannon1956.pdf`
[2] `http://www.wolframscience.com/prizes/tm23/`