Introduction to Information Retrieval
http://informationretrieval.org

IIR 4: Index Construction

Hinrich Schütze

Center for Information and Language Processing, University of Munich

2014-04-16
Dictionary as array of fixed-width entries

<table>
<thead>
<tr>
<th>term</th>
<th>document frequency</th>
<th>pointer to postings list</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>656,265</td>
<td>→</td>
</tr>
<tr>
<td>aachen</td>
<td>65</td>
<td>→</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>zulu</td>
<td>221</td>
<td>→</td>
</tr>
</tbody>
</table>

Space needed: 20 bytes 4 bytes 4 bytes
B-tree for looking up entries in array
Wildcard queries using a permuterm index
Wildcard queries using a permuterm index

Queries:
- For \( X \), look up \( X\$ \)
- For \( X^* \), look up \( X^*\$ \)
- For \( ^*X \), look up \( X^*\$ \)
- For \( ^*X^* \), look up \( X^* \)
- For \( X^*Y \), look up \( Y\$X^* \)
$k$-gram indexes for spelling correction: bordroom

- BO: aboard → about → boardroom → border
- OR: border → lord → morbid → sordid
- RD: aboard → ardent → boardroom → border
Levenshtein distance for spelling correction

\[
\text{LEVENSHTEIN DISTANCE}(s_1, s_2)
\]

1. for \( i \leftarrow 0 \) to \(|s_1|\) do \( m[i, 0] = i \)
2. for \( j \leftarrow 0 \) to \(|s_2|\) do \( m[0, j] = j \)
3. for \( i \leftarrow 1 \) to \(|s_1|\) do
4.   for \( j \leftarrow 1 \) to \(|s_2|\) do
5.     if \( s_1[i] = s_2[j] \) then \( m[i, j] = \min\{m[i - 1, j] + 1, m[i, j - 1] + 1, m[i - 1, j - 1]\} \)
6.     else \( m[i, j] = \min\{m[i - 1, j] + 1, m[i, j - 1] + 1, m[i - 1, j - 1] + 1\} \)
7. return \( m[|s_1|, |s_2|] \)

Operations: insert, delete, replace, copy
Exercise: Understand Peter Norvig’s spelling corrector

```python
import re, collections

def words(text):
    return re.findall('[a-z]+', text.lower())

def train(features):
    model = collections.defaultdict(lambda: 1)
    for f in features:
        model[f] += 1
    return model

NWORDS = train(words(file('big.txt').read()))

alphabet = 'abcdefghijklmnopqrstuvwxyz'

def edits1(word):
    splits = [(word[:i], word[i:]) for i in range(len(word) + 1)]
    deletes = [a + b[1:] for a, b in splits if b]
    transposes = [a + b[1] + b[0] + b[2:] for a, b in splits if len(b) > 1]
    replaces = [a + c + b[1:] for a, b in splits for c in alphabet if b]
    inserts = [a + c + b for a, b in splits for c in alphabet]
    return set(deletes + transposes + replaces + inserts)

def known_edits2(word):
    return set(e2 for e1 in edits1(word) for e2 in
               edits1(e1) if e2 in NWORDS)

def known(words):
    return set(w for w in words if w in NWORDS)

def correct(word):
    candidates = known([word]) or known(edits1(word)) or
    known_edits2(word) or [word]
    return max(candidates, key=NWORDS.get)
```
Take-away
Take-away

- Two index construction algorithms: **BSBI** (simple) and **SPIMI** (more realistic)
Take-away

- Two index construction algorithms: BSBI (simple) and SPIMI (more realistic)
- Distributed index construction: MapReduce
Take-away

- Two index construction algorithms: BSBI (simple) and SPIMI (more realistic)
- Distributed index construction: MapReduce
- Dynamic index construction: how to keep the index up-to-date as the collection changes
## Hardware basics
Many design decisions in information retrieval are based on hardware constraints.
Many design decisions in information retrieval are based on hardware constraints. We begin by reviewing hardware basics that we’ll need in this course.
Hardware basics
Access to data is much faster in memory than on disk. (roughly a factor of 10)
Hardware basics

- Access to data is much faster in memory than on disk. (roughly a factor of 10)
- Disk seeks are “idle” time: No data is transferred from disk while the disk head is being positioned.
Hardware basics

- Access to data is much **faster in memory than on disk.** (roughly a factor of 10)
- **Disk seeks are “idle” time:** No data is transferred from disk while the disk head is being positioned.
- To optimize transfer time from disk to memory: **one large chunk is faster than many small chunks.**
Access to data is much faster in memory than on disk. (roughly a factor of 10)

Disk seeks are “idle” time: No data is transferred from disk while the disk head is being positioned.

To optimize transfer time from disk to memory: one large chunk is faster than many small chunks.

Disk I/O is block-based: Reading and writing of entire blocks (as opposed to smaller chunks). Block sizes: 8KB to 256 KB
Hardware basics

- Access to data is much faster in memory than on disk. (roughly a factor of 10)
- Disk seeks are "idle" time: No data is transferred from disk while the disk head is being positioned.
- To optimize transfer time from disk to memory: one large chunk is faster than many small chunks.
- Disk I/O is block-based: Reading and writing of entire blocks (as opposed to smaller chunks). Block sizes: 8KB to 256 KB
- Servers used in IR systems typically have many GBs of main memory and TBs of disk space.
Access to data is much faster in memory than on disk. (roughly a factor of 10)

Disk seeks are "idle" time: No data is transferred from disk while the disk head is being positioned.

To optimize transfer time from disk to memory: one large chunk is faster than many small chunks.

Disk I/O is block-based: Reading and writing of entire blocks (as opposed to smaller chunks). Block sizes: 8KB to 256 KB

Servers used in IR systems typically have many GBs of main memory and TBs of disk space.

Fault tolerance is expensive: It’s cheaper to use many regular machines than one fault tolerant machine.
### Some stats (ca. 2008)

<table>
<thead>
<tr>
<th>symbol</th>
<th>statistic</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s$</td>
<td>average seek time</td>
<td>$5 \text{ ms} = 5 \times 10^{-3} \text{ s}$</td>
</tr>
<tr>
<td>$b$</td>
<td>transfer time per byte</td>
<td>$0.02 \mu \text{s} = 2 \times 10^{-8} \text{ s}$</td>
</tr>
<tr>
<td></td>
<td>processor’s clock rate</td>
<td>$10^9 \text{ s}^{-1}$</td>
</tr>
<tr>
<td>$p$</td>
<td>lowlevel operation (e.g., compare &amp; swap a word)</td>
<td>$0.01 \mu \text{s} = 10^{-8} \text{ s}$</td>
</tr>
<tr>
<td></td>
<td>size of main memory</td>
<td>several GB</td>
</tr>
<tr>
<td></td>
<td>size of disk space</td>
<td>$1 \text{ TB or more}$</td>
</tr>
</tbody>
</table>
RCV1 collection
Shakespeare’s collected works are not large enough for demonstrating many of the points in this course.
Shakespeare’s collected works are not large enough for demonstrating many of the points in this course.

As an example for applying scalable index construction algorithms, we will use the Reuters RCV1 collection.
Shakespeare’s collected works are not large enough for demonstrating many of the points in this course.

As an example for applying scalable index construction algorithms, we will use the Reuters RCV1 collection.

English newswire articles sent over the wire in 1995 and 1996 (one year).
Extreme conditions create rare Antarctic clouds

SYDNEY (Reuters) - Rare, mother-of-pearl colored clouds caused by extreme weather conditions above Antarctica are a possible indication of global warming, Australian scientists said on Tuesday.

Known as nacreous clouds, the spectacular formations showing delicate wisps of colors were photographed in the sky over an Australian
### Reuters RCV1 statistics

<table>
<thead>
<tr>
<th>$N$</th>
<th>documents</th>
<th>800,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>tokens per document</td>
<td>200</td>
</tr>
<tr>
<td>$M$</td>
<td>terms (≡ word types)</td>
<td>400,000</td>
</tr>
<tr>
<td></td>
<td>bytes per token (incl. spaces/punct.)</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>bytes per token (without spaces/punct.)</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>bytes per term (≡ word type)</td>
<td>7.5</td>
</tr>
<tr>
<td>$T$</td>
<td>non-positional postings</td>
<td>100,000,000</td>
</tr>
</tbody>
</table>
### Reuters RCV1 statistics

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>documents</td>
<td>800,000</td>
</tr>
<tr>
<td>$L$</td>
<td>tokens per document</td>
<td>200</td>
</tr>
<tr>
<td>$M$</td>
<td>terms (= word types)</td>
<td>400,000</td>
</tr>
<tr>
<td></td>
<td>bytes per token (incl. spaces/punct.)</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>bytes per token (without spaces/punct.)</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>bytes per term (= word type)</td>
<td>7.5</td>
</tr>
<tr>
<td>$T$</td>
<td>non-positional postings</td>
<td>100,000,000</td>
</tr>
</tbody>
</table>

**Exercise:** Average frequency of a term (how many tokens)? 4.5 bytes per word token vs. 7.5 bytes per word type: why the difference? How many positional postings?
Exercise

Why does this algorithm not scale to very large collections?
Goal: construct the inverted index

Brutus → 1 2 4 11 31 45 173 174
Caesar → 1 2 4 5 6 16 57 132 ...
Calpurnia → 2 31 54 101

: dictionary

: postings
### Index construction in IIR 1: Sort postings in memory

<table>
<thead>
<tr>
<th>term</th>
<th>docID</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>1</td>
</tr>
<tr>
<td>did</td>
<td>1</td>
</tr>
<tr>
<td>enact</td>
<td>1</td>
</tr>
<tr>
<td>julius</td>
<td>1</td>
</tr>
<tr>
<td>caesar</td>
<td>1</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>was</td>
<td>1</td>
</tr>
<tr>
<td>killed</td>
<td>1</td>
</tr>
<tr>
<td>i'</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
</tr>
<tr>
<td>capitol</td>
<td>1</td>
</tr>
<tr>
<td>brutus</td>
<td>1</td>
</tr>
<tr>
<td>killed</td>
<td>1</td>
</tr>
<tr>
<td>me</td>
<td>1</td>
</tr>
<tr>
<td>so</td>
<td>2</td>
</tr>
<tr>
<td>let</td>
<td>2</td>
</tr>
<tr>
<td>it</td>
<td>2</td>
</tr>
<tr>
<td>be</td>
<td>2</td>
</tr>
<tr>
<td>with</td>
<td>2</td>
</tr>
<tr>
<td>caesar</td>
<td>2</td>
</tr>
<tr>
<td>the</td>
<td>2</td>
</tr>
<tr>
<td>noble</td>
<td>2</td>
</tr>
<tr>
<td>brutus</td>
<td>2</td>
</tr>
<tr>
<td>hath</td>
<td>2</td>
</tr>
<tr>
<td>told</td>
<td>2</td>
</tr>
<tr>
<td>you</td>
<td>2</td>
</tr>
<tr>
<td>caesar</td>
<td>2</td>
</tr>
<tr>
<td>was</td>
<td>2</td>
</tr>
<tr>
<td>ambitious</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>term</th>
<th>docID</th>
</tr>
</thead>
<tbody>
<tr>
<td>ambitious</td>
<td>2</td>
</tr>
<tr>
<td>be</td>
<td>2</td>
</tr>
<tr>
<td>brutus</td>
<td>1</td>
</tr>
<tr>
<td>brutus</td>
<td>2</td>
</tr>
<tr>
<td>capitol</td>
<td>1</td>
</tr>
<tr>
<td>caesar</td>
<td>1</td>
</tr>
<tr>
<td>caesar</td>
<td>2</td>
</tr>
<tr>
<td>did</td>
<td>1</td>
</tr>
<tr>
<td>enact</td>
<td>1</td>
</tr>
<tr>
<td>hath</td>
<td>1</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>i'</td>
<td>1</td>
</tr>
<tr>
<td>it</td>
<td>2</td>
</tr>
<tr>
<td>julius</td>
<td>1</td>
</tr>
<tr>
<td>killed</td>
<td>1</td>
</tr>
<tr>
<td>killed</td>
<td>1</td>
</tr>
<tr>
<td>let</td>
<td>2</td>
</tr>
<tr>
<td>l</td>
<td>1</td>
</tr>
<tr>
<td>me</td>
<td>1</td>
</tr>
<tr>
<td>noble</td>
<td>2</td>
</tr>
<tr>
<td>so</td>
<td>2</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>2</td>
</tr>
<tr>
<td>told</td>
<td>2</td>
</tr>
<tr>
<td>told</td>
<td>2</td>
</tr>
<tr>
<td>you</td>
<td>2</td>
</tr>
<tr>
<td>was</td>
<td>1</td>
</tr>
<tr>
<td>was</td>
<td>2</td>
</tr>
<tr>
<td>with</td>
<td>2</td>
</tr>
</tbody>
</table>
Sort-based index construction
Sort-based index construction

- As we build index, we parse docs one at a time.
Sort-based index construction

- As we build index, we parse docs one at a time.
- The final postings for any term are incomplete until the end.
Sort-based index construction

- As we build index, we parse docs one at a time.
- The final postings for any term are incomplete until the end.
- Can we keep all postings in memory and then do the sort in-memory at the end?
Sort-based index construction

- As we build index, we parse docs one at a time.
- The final postings for any term are incomplete until the end.
- Can we keep all postings in memory and then do the sort in-memory at the end?
- No, not for large collections
Sort-based index construction

- As we build index, we parse docs one at a time.
- The final postings for any term are incomplete until the end.
- Can we keep all postings in memory and then do the sort in-memory at the end?
- No, not for large collections
- Thus: We need to store intermediate results on disk.
Same algorithm for disk?
Can we use the same index construction algorithm for larger collections, but by using disk instead of memory?
Can we use the same index construction algorithm for larger collections, but by using disk instead of memory?

No: Sorting very large sets of records on disk is too slow – too many disk seeks.
Can we use the same index construction algorithm for larger collections, but by using disk instead of memory?

No: Sorting very large sets of records on disk is too slow – too many disk seeks.

We need an external sorting algorithm.
“External” sorting algorithm (using few disk seeks)
“External” sorting algorithm (using few disk seeks)

- We must sort $T = 100,000,000$ non-positional postings.
"External" sorting algorithm (using few disk seeks)

- We must sort $T = 100,000,000$ non-positional postings.
  - Each posting has size 12 bytes (4+4+4: termID, docID, term frequency).
“External” sorting algorithm (using few disk seeks)

- We must sort \( T = 100,000,000 \) non-positional postings.
  - Each posting has size 12 bytes (4+4+4: termID, docID, term frequency).
- Define a block to consist of 10,000,000 such postings.
“External” sorting algorithm (using few disk seeks)

- We must sort $T = 100,000,000$ non-positional postings.
  - Each posting has size 12 bytes (4+4+4: termID, docID, term frequency).
- Define a block to consist of 10,000,000 such postings
  - We can easily fit that many postings into memory.
“External” sorting algorithm (using few disk seeks)

- We must sort $T = 100,000,000$ non-positional postings.
  - Each posting has size 12 bytes (4+4+4: termID, docID, term frequency).
- Define a block to consist of 10,000,000 such postings
  - We can easily fit that many postings into memory.
  - We will have 10 such blocks for RCV1.
We must sort $T = 100,000,000$ non-positional postings.
- Each posting has size 12 bytes (4+4+4: termID, docID, term frequency).
- Define a block to consist of 10,000,000 such postings
  - We can easily fit that many postings into memory.
  - We will have 10 such blocks for RCV1.

Basic idea of algorithm:
“External” sorting algorithm (using few disk seeks)

- We must sort $T = 100,000,000$ non-positional postings.
  - Each posting has size 12 bytes (4+4+4: termID, docID, term frequency).
- Define a block to consist of 10,000,000 such postings
  - We can easily fit that many postings into memory.
  - We will have 10 such blocks for RCV1.
- Basic idea of algorithm:
  - For each block: (i) accumulate postings, (ii) sort in memory, (iii) write to disk
“External” sorting algorithm (using few disk seeks)

- We must sort $T = 100,000,000$ non-positional postings.
  - Each posting has size 12 bytes (4+4+4: termID, docID, term frequency).
- Define a block to consist of 10,000,000 such postings
  - We can easily fit that many postings into memory.
  - We will have 10 such blocks for RCV1.
- Basic idea of algorithm:
  - For each block: (i) accumulate postings, (ii) sort in memory, (iii) write to disk
  - Then merge the blocks into one long sorted order.
Merging two blocks

postings to be merged

<table>
<thead>
<tr>
<th>Block 1</th>
<th>Block 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>brutus</td>
<td>brutus</td>
</tr>
<tr>
<td>d3</td>
<td>d2</td>
</tr>
<tr>
<td>caesar</td>
<td>caesar</td>
</tr>
<tr>
<td>d4</td>
<td>d1</td>
</tr>
<tr>
<td>noble</td>
<td>julius</td>
</tr>
<tr>
<td>d3</td>
<td>d1</td>
</tr>
<tr>
<td>with</td>
<td>killed</td>
</tr>
<tr>
<td>d4</td>
<td>d2</td>
</tr>
</tbody>
</table>

merged postings

<table>
<thead>
<tr>
<th>brutus</th>
<th>brutus</th>
</tr>
</thead>
<tbody>
<tr>
<td>d2</td>
<td>d3</td>
</tr>
<tr>
<td>caesar</td>
<td>caesar</td>
</tr>
<tr>
<td>d1</td>
<td>d4</td>
</tr>
<tr>
<td>julius</td>
<td>killed</td>
</tr>
<tr>
<td>d1</td>
<td>d2</td>
</tr>
<tr>
<td>noble</td>
<td>noble</td>
</tr>
<tr>
<td>d3</td>
<td>d3</td>
</tr>
<tr>
<td>with</td>
<td>with</td>
</tr>
<tr>
<td>d4</td>
<td>d4</td>
</tr>
</tbody>
</table>

disk
Blocked Sort-Based Indexing

```
BSBIndexConstruction()
1   n ← 0
2   while (all documents have not been processed)
3     do n ← n + 1
4      block ← ParseNextBlock()
5      BSBI-Invert(block)
6      WriteBlockToDisk(block, f_n)
7      MergeBlocks(f_1, ..., f_n; f_{merged})
```
Problem with sort-based algorithm
Problem with sort-based algorithm

- Our assumption was: we can keep the dictionary in memory.
Problem with sort-based algorithm

- Our assumption was: we can keep the dictionary in memory.
- We need the dictionary (which grows dynamically) in order to implement a term to termID mapping.
Problem with sort-based algorithm

- Our assumption was: we can keep the dictionary in memory.
- We need the dictionary (which grows dynamically) in order to implement a term to termID mapping.
- Actually, we could work with term,docID postings instead of termID,docID postings . . .
Problem with sort-based algorithm

- Our assumption was: we can keep the dictionary in memory.
- We need the dictionary (which grows dynamically) in order to implement a term to termID mapping.
- Actually, we could work with term,docID postings instead of termID,docID postings ...
- ... but then intermediate files become very large. (We would end up with a scalable, but very slow index construction method.)
Single-pass in-memory indexing
Single-pass in-memory indexing

- Abbreviation: SPIMI
Single-pass in-memory indexing

- Abbreviation: SPIMI
- Key idea 1: Generate separate dictionaries for each block – no need to maintain term-termID mapping across blocks.
Abbreviation: SPIMI

Key idea 1: Generate separate dictionaries for each block – no need to maintain term-termID mapping across blocks.

Key idea 2: Don’t sort. Accumulate postings in postings lists as they occur.
Single-pass in-memory indexing

- Abbreviation: SPIMI
- Key idea 1: Generate separate dictionaries for each block – no need to maintain term-termID mapping across blocks.
- Key idea 2: Don’t sort. Accumulate postings in postings lists as they occur.
- With these two ideas we can generate a complete inverted index for each block.
Single-pass in-memory indexing

- Abbreviation: SPIMI
- Key idea 1: Generate separate dictionaries for each block – no need to maintain term-termID mapping across blocks.
- Key idea 2: Don’t sort. Accumulate postings in postings lists as they occur.
- With these two ideas we can generate a complete inverted index for each block.
- These separate indexes can then be merged into one big index.
**SPIMI-Invert**

\[
\text{SPIMI-Invert}(\text{token\_stream})
\]

1. \[ \text{output\_file} \leftarrow \text{NEWFILE}() \]
2. \[ \text{dictionary} \leftarrow \text{NEWHASH}() \]
3. \[ \text{while} \ (\text{free memory available}) \]
4. \[ \text{do} \ \text{token} \leftarrow \text{next}(\text{token\_stream}) \]
5. \[ \text{if} \ \text{term}(\text{token}) \notin \text{dictionary} \]
6. \[ \text{then} \ \text{postings\_list} \leftarrow \text{ADDTO DICTIONARY}(\text{dictionary}, \text{term}(\text{token})) \]
7. \[ \text{else} \ \text{postings\_list} \leftarrow \text{GETPOSTINGSLIST}(\text{dictionary}, \text{term}(\text{token})) \]
8. \[ \text{if} \ \text{full}(\text{postings\_list}) \]
9. \[ \text{then} \ \text{postings\_list} \leftarrow \text{DOUBLEPOSTINGSLIST}(\text{dictionary}, \text{term}(\text{token})) \]
10. \[ \text{ADDTO POSTINGSLIST}(\text{postings\_list}, \text{docID}(\text{token})) \]
11. \[ \text{sorted\_terms} \leftarrow \text{SORT TERMS}(\text{dictionary}) \]
12. \[ \text{WRITE BLOCK TO DISK}(\text{sorted\_terms}, \text{dictionary}, \text{output\_file}) \]
13. \[ \text{return} \ \text{output\_file} \]
SPIMI-Invert

**SPIMI-Invert** *(token_stream)*

1. `output_file ← NewFile()`
2. `dictionary ← NewHash()`
3. while (free memory available)
4. do token ← `next(token_stream)`
5. if `term(token) ∉ dictionary`
6. then `postings_list ← AddToDictionary(dictionary, term(token))`
7. else `postings_list ← GetPostingsList(dictionary, term(token))`
8. if `full(postings_list)`
9. then `postings_list ← DoublePostingsList(dictionary, term(token))`
10. `AddToPostingsList(postings_list, docID(token))`
11. `sorted_terms ← SortTerms(dictionary)`
12. `WriteBlockToDisk(sorted_terms, dictionary, output_file)`
13. return `output_file`

Merging of blocks is analogous to BSBI.
<table>
<thead>
<tr>
<th>Recap</th>
<th>Introduction</th>
<th>BSBI algorithm</th>
<th><strong>SPIMI algorithm</strong></th>
<th>Distributed indexing</th>
<th>Dynamic indexing</th>
</tr>
</thead>
</table>

**SPIMI: Compression**
**SPIMI: Compression**

- Compression makes SPIMI even more efficient.
SPIMI: Compression

- Compression makes SPIMI even more efficient.
  - Compression of terms
SPIMI: Compression

- Compression makes SPIMI even more efficient.
  - Compression of terms
  - Compression of postings
Compression makes SPIMI even more efficient.

- Compression of terms
- Compression of postings
- See next lecture
Outline

1 Recap
2 Introduction
3 BSBI algorithm
4 SPIMI algorithm
5 Distributed indexing
6 Dynamic indexing
Distributed indexing
Distributed indexing

- For web-scale indexing (don't try this at home!): must use a distributed computer cluster
Distributed indexing

- For web-scale indexing (don’t try this at home!): must use a distributed computer cluster
- Individual machines are fault-prone.
Distributed indexing

- For web-scale indexing (don’t try this at home!): must use a distributed computer cluster
- Individual machines are fault-prone.
  - Can unpredictably slow down or fail.
For web-scale indexing (don’t try this at home!): must use a distributed computer cluster

- Individual machines are fault-prone.
  - Can unpredictably slow down or fail.

- How do we exploit such a pool of machines?
Google data centers (2007 estimates; Gartner)
Google data centers (2007 estimates; Gartner)

- Google data centers mainly contain commodity machines.
Google data centers (2007 estimates; Gartner)

- Google data centers mainly contain commodity machines.
- Data centers are distributed all over the world.
Google data centers (2007 estimates; Gartner)

- Google data centers mainly contain commodity machines.
- Data centers are distributed all over the world.
- 1 million servers, 3 million processors/cores
Google data centers (2007 estimates; Gartner)

- Google data centers mainly contain commodity machines.
- Data centers are distributed all over the world.
- 1 million servers, 3 million processors/cores
- Google installs 100,000 servers each quarter.
Google data centers (2007 estimates; Gartner)

- Google data centers mainly contain commodity machines.
- Data centers are distributed all over the world.
- 1 million servers, 3 million processors/cores
- Google installs 100,000 servers each quarter.
- Based on expenditures of 200–250 million dollars per year
Google data centers (2007 estimates; Gartner)

- Google data centers mainly contain commodity machines.
- Data centers are distributed all over the world.
- 1 million servers, 3 million processors/cores
- Google installs 100,000 servers each quarter.
- Based on expenditures of 200–250 million dollars per year
- This would be 10% of the computing capacity of the world!
Google data centers (2007 estimates; Gartner)

- Google data centers mainly contain commodity machines.
- Data centers are distributed all over the world.
- 1 million servers, 3 million processors/cores
- Google installs 100,000 servers each quarter.
- Based on expenditures of 200–250 million dollars per year
- This would be 10% of the computing capacity of the world!
- If in a non-fault-tolerant system with 1000 nodes, each node has 99.9% uptime, what is the uptime of the system (assuming it does not tolerate failures)?
Google data centers (2007 estimates; Gartner)

- Google data centers mainly contain commodity machines.
- Data centers are distributed all over the world.
- 1 million servers, 3 million processors/cores
- Google installs 100,000 servers each quarter.
- Based on expenditures of 200–250 million dollars per year
- This would be 10% of the computing capacity of the world!
- If in a non-fault-tolerant system with 1000 nodes, each node has 99.9% uptime, what is the uptime of the system (assuming it does not tolerate failures)?
- Answer: 37%
Google data centers (2007 estimates; Gartner)

- Google data centers mainly contain commodity machines.
- Data centers are distributed all over the world.
- 1 million servers, 3 million processors/cores
- Google installs 100,000 servers each quarter.
- Based on expenditures of 200–250 million dollars per year
- This would be 10% of the computing capacity of the world!
- If in a non-fault-tolerant system with 1000 nodes, each node has 99.9% uptime, what is the uptime of the system (assuming it does not tolerate failures)?
  - Answer: 37%
- Suppose a server will fail after 3 years. For an installation of 1 million servers, what is the interval between machine failures?
Google data centers (2007 estimates; Gartner)

- Google data centers mainly contain commodity machines.
- Data centers are distributed all over the world.
- 1 million servers, 3 million processors/cores
- Google installs 100,000 servers each quarter.
- Based on expenditures of 200–250 million dollars per year
- This would be 10% of the computing capacity of the world!

If in a non-fault-tolerant system with 1000 nodes, each node has 99.9% uptime, what is the uptime of the system (assuming it does not tolerate failures)?

- Answer: 37%

Suppose a server will fail after 3 years. For an installation of 1 million servers, what is the interval between machine failures?

- Answer: less than two minutes
Distributed indexing
Distributed indexing

- Maintain a master machine directing the indexing job – considered “safe”
Distributed indexing

- Maintain a **master** machine directing the indexing job – considered “safe”
- Break up indexing into sets of parallel tasks
Distributed indexing

- Maintain a master machine directing the indexing job – considered “safe”
- Break up indexing into sets of parallel tasks
- Master machine assigns each task to an idle machine from a pool.
Parallel tasks
Parallel tasks

- We will define two sets of parallel tasks and deploy two types of machines to solve them:
Parallel tasks

- We will define two sets of parallel tasks and deploy two types of machines to solve them:
  - Parsers
Parallel tasks

- We will define two sets of parallel tasks and deploy two types of machines to solve them:
  - Parsers
  - Inverters
Parallel tasks

- We will define two sets of parallel tasks and deploy two types of machines to solve them:
  - Parsers
  - Inverters

- Break the input document collection into \textit{splits} (corresponding to blocks in BSBI/SPIMI)
Parallel tasks

- We will define two sets of parallel tasks and deploy two types of machines to solve them:
  - Parsers
  - Inverters
- Break the input document collection into *splits* (corresponding to blocks in BSBI/SPIMI)
- Each split is a subset of documents.
Parsers

- Master assigns a split to an idle parser machine.
- Parser reads a document at a time and emits (term, docID)-pairs.
- Parser writes pairs into $j$ term-partitions.
- Each for a range of terms’ first letters
  - E.g., a-f, g-p, q-z (here: $j = 3$)
Inverters

- An inverter collects all (term,docID) pairs (= postings) for one term-partition (e.g., for a-f).
- Sorts and writes to postings lists
Data flow

- **splits**
- **map phase**
- **segment files**
- **reduce phase**

- **assign**
- **master**
- **assign**

- **parser**
  - a-f g-p q-z

- **inverter**
  - a-f
  - g-p
  - q-z

- **postings**
MapReduce
The index construction algorithm we just described is an instance of MapReduce.
MapReduce

- The index construction algorithm we just described is an instance of MapReduce.
- MapReduce is a robust and conceptually simple framework for distributed computing . . .
MapReduce

- The index construction algorithm we just described is an instance of MapReduce.
- MapReduce is a robust and conceptually simple framework for distributed computing . . .
- . . . without having to write code for the distribution part.
MapReduce

- The index construction algorithm we just described is an instance of MapReduce.
- MapReduce is a robust and conceptually simple framework for distributed computing . . .
- . . . without having to write code for the distribution part.
- The Google indexing system (ca. 2002) consisted of a number of phases, each implemented in MapReduce.
The index construction algorithm we just described is an instance of MapReduce.

MapReduce is a robust and conceptually simple framework for distributed computing . . .

. . . without having to write code for the distribution part.

The Google indexing system (ca. 2002) consisted of a number of phases, each implemented in MapReduce.

Index construction was just one phase.
The index construction algorithm we just described is an instance of MapReduce.

MapReduce is a robust and conceptually simple framework for distributed computing . . .

. . . without having to write code for the distribution part.

The Google indexing system (ca. 2002) consisted of a number of phases, each implemented in MapReduce.

Index construction was just one phase.

Another phase: transform term-partitioned into document-partitioned index.
MapReduce

- The index construction algorithm we just described is an instance of MapReduce.
- MapReduce is a robust and conceptually simple framework for distributed computing . . .
- . . . without having to write code for the distribution part.
- The Google indexing system (ca. 2002) consisted of a number of phases, each implemented in MapReduce.
- Index construction was just one phase.
- Another phase: transform term-partitioned into document-partitioned index.
- Why might a document-partitioned index be preferable?
Index construction in MapReduce

Schema of map and reduce functions

map: input → list(k, v)
reduce: (k, list(v)) → output

Instantiation of the schema for index construction

map: web collection → list(termID, docID)
reduce: (⟨termID1, list(docID)⟩, ⟨termID2, list(docID)⟩, . . .) → (postings_list1, postings_list2, . . .)

Example for index construction

map: d2 : C DIED. d1 : C CAME, C c’ED. → ((C, d2), ⟨DIED, d2⟩, ⟨C, d1⟩, ⟨CAME, d1⟩, ⟨C, d1⟩, ⟨c’ED, d1⟩)
reduce: ((C, ⟨d2, d1⟩), ⟨DIED, ⟨d2⟩⟩, ⟨CAME, ⟨d1⟩⟩, ⟨c’ED, ⟨d1⟩⟩) → ((C, ⟨d1:2, d2:1⟩), ⟨DIED, ⟨d2:1⟩⟩, ⟨CAME, ⟨d1:1⟩⟩, ⟨c’ED, ⟨d1:1⟩⟩)
Exercise

- What information does the task description contain that the master gives to a parser?
- What information does the parser report back to the master upon completion of the task?
- What information does the task description contain that the master gives to an inverter?
- What information does the inverter report back to the master upon completion of the task?
Outline

1 Recap
2 Introduction
3 BSBI algorithm
4 SPIMI algorithm
5 Distributed indexing
6 Dynamic indexing
Dynamic indexing
Dynamic indexing

- Up to now, we have assumed that collections are static.
Dynamic indexing

- Up to now, we have assumed that collections are **static**.
- They rarely are: Documents are inserted, deleted and modified.
Dynamic indexing

- Up to now, we have assumed that collections are **static**.
- They rarely are: Documents are inserted, deleted and modified.
- This means that the dictionary and postings lists have to be **dynamically** modified.
Dynamic indexing: Simplest approach
Dynamic indexing: Simplest approach

- Maintain big main index on disk
Dynamic indexing: Simplest approach

- Maintain big main index on disk
- New docs go into small auxiliary index in memory.
Dynamic indexing: Simplest approach

- Maintain **big main index on disk**
- New docs go into **small auxiliary index in memory**.
- Search across both, merge results
Dynamic indexing: Simplest approach

- Maintain **big main index on disk**
- New docs go into **small auxiliary index in memory**.
- Search across both, merge results
- Periodically, merge auxiliary index into big index
Dynamic indexing: Simplest approach

- Maintain **big main index on disk**
- New docs go into **small auxiliary index in memory**.
- Search across both, merge results
- Periodically, merge auxiliary index into big index
- Deletions:
Dynamic indexing: Simplest approach

- Maintain **big main index on disk**
- New docs go into **small auxiliary index in memory**.
- Search across both, merge results
- Periodically, merge auxiliary index into big index
- Deletions:
  - Invalidation bit-vector for deleted docs
Dynamic indexing: Simplest approach

- Maintain big main index on disk
- New docs go into small auxiliary index in memory.
- Search across both, merge results
- Periodically, merge auxiliary index into big index
- Deletions:
  - Invalidation bit-vector for deleted docs
  - Filter docs returned by index using this bit-vector
Issue with auxiliary and main index
Issue with auxiliary and main index

- Frequent merges
Issue with auxiliary and main index

- Frequent merges
- Poor search performance during index merge
Logarithmic merge
Logarithmic merge

- Logarithmic merging amortizes the cost of merging indexes over time.
Logarithmic merge

- Logarithmic merging amortizes the cost of merging indexes over time.
  - Users see smaller effect on response times.
Logarithmic merge

- Logarithmic merging amortizes the cost of merging indexes over time.
  - Users see smaller effect on response times.
- Maintain a series of indexes, each twice as large as the previous one.
Logarithmic merge

- Logarithmic merging amortizes the cost of merging indexes over time.
  - Users see smaller effect on response times.
- Maintain a series of indexes, each twice as large as the previous one.
- Keep smallest ($Z_0$) in memory
Logarithmic merge

- Logarithmic merging amortizes the cost of merging indexes over time.
  - Users see smaller effect on response times.
- Maintain a series of indexes, each twice as large as the previous one.
- Keep smallest ($Z_0$) in memory
- Larger ones ($l_0, l_1, \ldots$) on disk
Logarithmic merge

- Logarithmic merging amortizes the cost of merging indexes over time.
  - \( \rightarrow \) Users see smaller effect on response times.
- Maintain a series of indexes, each twice as large as the previous one.
- Keep smallest \((Z_0)\) in memory
- Larger ones \((I_0, I_1, \ldots)\) on disk
- If \(Z_0\) gets too big \((> n)\), write to disk as \(I_0\)
Logarithmic merge

- Logarithmic merging amortizes the cost of merging indexes over time.
  - Users see smaller effect on response times.
- Maintain a series of indexes, each twice as large as the previous one.
- Keep smallest ($Z_0$) in memory
- Larger ones ($l_0$, $l_1$, ...) on disk
- If $Z_0$ gets too big ($> n$), write to disk as $l_0$
- ... or merge with $l_0$ (if $l_0$ already exists) and write merger to $l_1$ etc.
**LMergeAddToken** *(indexes, Z₀, token)*

1. \[ Z₀ \leftarrow \text{Merge}(Z₀, \{ \text{token} \}) \]
2. if \( |Z₀| = n \)
3. then for \( i \leftarrow 0 \) to \( \infty \)
4. do if \( I_i \in \text{indexes} \)
5. then \( Z_{i+1} \leftarrow \text{Merge}(I_i, Z_i) \)
6. \( (Z_{i+1} \text{ is a temporary index on disk.}) \)
7. \( \text{indexes} \leftarrow \text{indexes} - \{ I_i \} \)
8. else \( I_i \leftarrow Z_i \) \( (Z_i \text{ becomes the permanent index } I_i.) \)
9. \( \text{indexes} \leftarrow \text{indexes} \cup \{ I_i \} \)
10. **Break**
11. \( Z₀ \leftarrow \emptyset \)

**LogarithmicMerge()**

1. \( Z₀ \leftarrow \emptyset \) \( (Z₀ \text{ is the in-memory index.}) \)
2. \( \text{indexes} \leftarrow \emptyset \)
3. **while** true
4. do **LMergeAddToken** *(indexes, Z₀, getNextToken())*
Binary numbers: \( l_3 l_2 l_1 l_0 = 2^3 2^2 2^1 2^0 \)
Binary numbers: \[ I_3 I_2 I_1 I_0 = 2^3 2^2 2^1 2^0 \]

- 0001
Binary numbers: \( l_3l_2l_1l_0 = 2^32^22^12^0 \)

- 0001
- 0010
Binary numbers: $l_3l_2l_1l_0 = 2^32^22^12^0$

- 0001
- 0010
- 0011
Binary numbers: $l_3 l_2 l_1 l_0 = 2^3 2^2 2^1 2^0$

- 0001
- 0010
- 0011
- 0100
Binary numbers: $l_3l_2l_1l_0 = 2^32^22^12^0$

- 0001
- 0010
- 0011
- 0100
- 0101
Binary numbers: \( l_3 l_2 l_1 l_0 = 2^3 2^2 2^1 2^0 \)

- 0001
- 0010
- 0011
- 0100
- 0101
- 0110
Binary numbers: $l_3l_2l_1l_0 = 2^32^22^12^0$

- 0001
- 0010
- 0011
- 0100
- 0101
- 0110
- 0111
Binary numbers: $l_3l_2l_1l_0 = 2^32^22^12^0$

- 0001
- 0010
- 0011
- 0100
- 0101
- 0110
- 0111
- 1000
Binary numbers: $l_3l_2l_1l_0 = 2^3 2^2 2^1 2^0$

- 0001
- 0010
- 0011
- 0100
- 0101
- 0110
- 0111
- 1000
- 1001
Binary numbers: $l_3 l_2 l_1 l_0 = 2^3 2^2 2^1 2^0$

- 0001
- 0010
- 0011
- 0100
- 0101
- 0110
- 0111
- 1000
- 1001
- 1010
Binary numbers: $l_3l_2l_1l_0 = 2^32^22^12^0$

- 0001
- 0010
- 0011
- 0100
- 0101
- 0110
- 0111
- 1000
- 1001
- 1010
- 1011
Binary numbers: $l_3l_2l_1l_0 = 2^32^22^12^0$

- 0001
- 0010
- 0011
- 0100
- 0101
- 0110
- 0111
- 1000
- 1001
- 1010
- 1011
- 1100
Logarithmic merge
Logarithmic merge

- Number of indexes bounded by $O(\log T)$ ($T$ is total number of postings read so far)
Logarithmic merge

- Number of indexes bounded by $O(\log T)$ ($T$ is total number of postings read so far)
- So query processing requires the merging of $O(\log T)$ indexes
Logarithmic merge

- Number of indexes bounded by $O(\log T)$ ($T$ is total number of postings read so far)
- So query processing requires the merging of $O(\log T)$ indexes
- Time complexity of index construction is $O(T \log T)$. 
Logarithmic merge

- Number of indexes bounded by $O(\log T)$ ($T$ is total number of postings read so far)
- So query processing requires the merging of $O(\log T)$ indexes
- Time complexity of index construction is $O(T \log T)$.
  - ... because each of $T$ postings is merged $O(\log T)$ times.
Logarithmic merge

- Number of indexes bounded by $O(\log T)$ ($T$ is total number of postings read so far)
- So query processing requires the merging of $O(\log T)$ indexes
- Time complexity of index construction is $O(T \log T)$.
  - ... because each of $T$ postings is merged $O(\log T)$ times.
- Auxiliary index: index construction time is $O(T^2)$ as each posting is touched in each merge.
Logarithmic merge

- Number of indexes bounded by $O(\log T)$ ($T$ is total number of postings read so far)
- So query processing requires the merging of $O(\log T)$ indexes
- Time complexity of index construction is $O(T \log T)$.
  - ... because each of $T$ postings is merged $O(\log T)$ times.
- Auxiliary index: index construction time is $O(T^2)$ as each posting is touched in each merge.
  - Suppose auxiliary index has size $a$
Logarithmic merge

- Number of indexes bounded by $O(\log T)$ ($T$ is total number of postings read so far)
- So query processing requires the merging of $O(\log T)$ indexes
- Time complexity of index construction is $O(T \log T)$.
  - . . . because each of $T$ postings is merged $O(\log T)$ times.
- Auxiliary index: index construction time is $O(T^2)$ as each posting is touched in each merge.
  - Suppose auxiliary index has size $a$
  - $a + 2a + 3a + 4a + \ldots + na = a\frac{n(n+1)}{2} = O(n^2)$
Logarithmic merge

- Number of indexes bounded by $O(\log T)$ ($T$ is total number of postings read so far)
- So query processing requires the merging of $O(\log T)$ indexes
- Time complexity of index construction is $O(T \log T)$.
  - ... because each of $T$ postings is merged $O(\log T)$ times.
- Auxiliary index: index construction time is $O(T^2)$ as each posting is touched in each merge.
  - Suppose auxiliary index has size $a$
    - $a + 2a + 3a + 4a + \ldots + na = a\frac{n(n+1)}{2} = O(n^2)$
- So logarithmic merging is an order of magnitude more efficient.
Dynamic indexing at large search engines
Dynamic indexing at large search engines

- Often a combination
Dynamic indexing at large search engines

- Often a combination
  - Frequent incremental changes
Dynamic indexing at large search engines

- Often a combination
  - Frequent incremental changes
  - Rotation of large parts of the index that can then be swapped in
Often a combination

- Frequent incremental changes
- Rotation of large parts of the index that can then be swapped in
- Occasional complete rebuild (becomes harder with increasing size – not clear if Google can do a complete rebuild)
Building positional indexes
Building positional indexes

- Basically the same problem except that the intermediate data structures are large.
Take-away

- Two index construction algorithms: **BSBI** (simple) and **SPIMI** (more realistic)
- **Distributed** index construction: MapReduce
- **Dynamic** index construction: how to keep the index up-to-date as the collection changes
Chapter 4 of IIR

Resources at http://cislmu.org
  - Original publication on MapReduce by Dean and Ghemawat (2004)
  - Original publication on SPIMI by Heinz and Zobel (2003)
  - YouTube video: Google data centers