Introdução

Índices

Visão geral

Características gerais

Agrupamento

Denso × esparso

Primário

Secundário

Créditos

Organização e Recuperação da Informação

Indexação – Parte 1

Jander Moreira

UFSCar – DC

11 de setembro de 2017

Este material é complementar, sendo apenas uma apresentação de aula e não consistindo em material suficiente para estudo.

Agenda

1. Introdução
2. Índices
   - Visão geral
   - Características gerais
   - Agrupamento
   - Denso × esparso
   - Primário
   - Secundário

Material baseado em
FOLK, M.J & ZOELLICK, B. *File structures*, 2^{nd} ed.
Addison-Wesley Publishing Company, Inc. 1992

Capítulo 4, seção 4.2.2

Capítulo 6

Introdução

Índices
Visão geral
Características gerais
Agrupamento
Denso × esparso
Primário
Secundário

Créditos

Introdução

Introdução

Índice?

Imagem: Pixabay

Exemplos

- Lista de palavras-chaves no fim dos livros e as páginas onde ocorrem
- Localização de um título na biblioteca: piso, estante...
- ...

Objetivos

- Facilitar o acesso
- Aumentar o desempenho no acesso

Introdução

Índice

Imagem: Pixabay

Em arquivos

- Objetivo: recuperação de um (ou mais) registros
- Organização em um arquivo auxiliar contendo a estrutura do índice
- Uso de uma chave de pesquisa

Introdução

Blocos/páginas

Um bloco[1] é a unidade de transferência e de armazenamento de um sistema de arquivos.

- Transferências em unidades de 1 bloco
- Divisão de armazenamento em unidades de 1 bloco
- Objetivo: otimizar o tempo de acesso (∴ melhorar desempenho)

Imagem: Pixabay

[1] Ou *página*, ou *cluster*.

Introdução

Arquivos e blocos

A organização dos registros dos arquivos em blocos de disco é essencial para melhorar o desempenho de acesso.

Imagem: Pixabay

Introdução

Índices

Visão geral
Características gerais
Agrupamento
Denso × esparso
Primário
Secundário

Créditos

Índices

Índices

Dados para exemplos (baseado de Folk e Zoellick, 1992)

```
LON|2312|Romeo e Julieta|Prokofiev|Maazel|#
RCA|2626|Quarteto em C menor|Beethoven|Julliard|#
WAR|23699|Touchstone|Corea|Corea|#
ANG|3795|Sinfonia n. 9|Beethoven|Giulini|#
COL|38358|Nebraska|Springsteen|Springsteen|#
DG|18807|Sinfonia n. 9|Beethoven|Karajan|#
MER|75016|Coq d'Or Suite|Rimsky-Korsakov|Leinsdorf|#
COL|31809|Sinfonia n. 9|Dvorak|Bernstein|#
DG|139201|Concerto de Violino|Beethoven|Ferras|#
FF|245|Good News|Sweet Honey Rock|Sweet Honey|#
```

Esquema ilustrado com um registro por linha

Índices

Arquivo de dados

```
000 L O N | 2 3 1 2 | R o m e o _ e _ J u l i e t a | P r o k o
030 f i e v | M a a z e l | # R C A | 2 6 2 6 | Q u a r t e t o
060 _ e m _ C _ m e n o r | B e e t h o v e n | J u l l i a r d
090 | # W A R | 2 3 6 9 9 | T o u c h s t o n e | C o r e a | C
120 o r e a | # A N G | 3 7 9 5 | S i n f o n i a _ n . _ 9 | B
150 e e t h o v e n | G i u l i n i | # C O L | 3 8 3 5 8 | N e
180 b r a s k a | S p r i n g s t e e n | S p r i n g s t e e n
210 | # D G | 1 8 8 0 7 | S i n f o n i a _ n . _ 9 | B e e t h
240 o v e n | K a r a j a n | # M E R | 7 5 0 1 6 | C o q _ d '
270 O r _ S u i t e | R i m s k y - K o r s a k o v | L e i n s
300 d o r f | # C O L | 3 1 8 0 9 | S i n f o n i a _ n . _ 9 |
330 D v o r a k | B e r n s t e i n | # D G | 1 3 9 2 0 1 | C o
360 n c e r t o _ d e _ V i o l i n o | B e e t h o v e n | F e
390 r r a s | # F F | 2 4 5 | G o o d _ N e w s | S w e e t _ H
420 o n e y _ R o c k | S w e e t _ H o n e y | #
```

Registros e campos de tamanho variável, com delimitadores de campo (|) e de registro (#)

Campos: `gravadora` (sigla), `ID`, `título`, `compositor` e `artista`

Índices

Índice pela chave `gravadora+ID`

Chave	**Endereço** (*byte offset*)
ANG3795	126
COL31809	306
COL38358	168
DG139201	348
DG18807	212
FF245	396
LON2312	0
MER75016	254
RCA2626	43
WAR23699	92

Índices

Índice pela chave `compositor`

Chave	**Endereço** (*byte offset*)
Beethoven	43
Beethoven	126
Beethoven	212
Beethoven	348
Corea	92
Dvorak	306
Prokofiev	0
Rimsky-Korsakov	254
Springsteen	168
Sweet Honey Rock	396

Exercício I

Construa o índice pela chave `artista`.

Arquivo de dados (visão por registros)

```
000 LON|2312|Romeo e Julieta|Prokofiev|Maazel|#
043 RCA|2626|Quarteto em C menor|Beethoven|Julliard|#
092 WAR|23699|Touchstone|Corea|Corea|#
126 ANG|3795|Sinfonia n. 9|Beethoven|Giulini|#
168 COL|38358|Nebraska|Springsteen|Springsteen|#
212 DG|18807|Sinfonia n. 9|Beethoven|Karajan|#
254 MER|75016|Coq d'Or Suite|Rimsky-Korsakov|Leinsdorf|#
306 COL|31809|Sinfonia n. 9|Dvorak|Bernstein|#
348 DG|139201|Concerto de Violino|Beethoven|Ferras|#
396 FF|245|Good News|Sweet Honey Rock|Sweet Honey|#
```

Exercício II

Resposta – Índice pela chave `compositor`

Chave	**Endereço** (*byte offset*)
Bernstein	306
Corea	92
Ferras	348
Giulini	126
Julliard	43
Karajan	212
Leinsdorf	254
Maazel	0
Springsteen	168
Sweet Honey	396

Índices

Eficiência da busca com índice

- Ordenado pela chave de busca
- Entradas de índice menores que registros de dados
  - Arquivo de índice com menos bytes
  - Menos acesso a disco
  - Acesso direto ao arquivo de dados
- Possibilidade de manutenção em memória principal
  - Busca no índice sem custo de acesso a disco
  - Acesso a disco na recuperação do registro

Introdução

Índices

Visão geral

Características gerais

Agrupamento

Denso × esparso

Primário

Secundário

Créditos

Índices

Agrupado × não agrupado

Índice agrupado

O índice é dito agrupado quando a ordem dos registros de dados coincide com a ordem das entradas de índice

Índice não agrupado

Um índice não agrupado não mantém controle entre a ordem das entradas de índice e a ordem dos registros de dados

Índices

Exemplo: agrupado

Índice (gravadora+ID)

Chave	Endereço
ANG3795	0
COL31809	42
COL38358	84
DG139201	128
DG18807	176
FF245	218
LON2312	265
MER75016	308
RCA2626	360
WAR23699	408

Arquivo de dados

```
000 ANG|3795|Sinfonia_n._9|Beethoven|Giu
036 lini|#COL|31809|Sinfonia_n._9|Dvorak
072 |Bernstein|#COL|38358|Nebraska|Sprin
108 gsteen|Springsteen|#DG|139201|Concer
144 to_de_Violino|Beethoven|Ferras|#DG|1
180 8807|Sinfonia_n._9|Beethoven|Karajan
216 |#FF|245|Good_News|Sweet_Honey_Rock|
252 Sweet_Honey|#LON|2312|Romeo_e_Juliet
288 a|Prokofiev|Maazel|#MER|75016|Coq_d'
324 Or_Suite|Rimsky-Korsakov|Leinsdorf|#
360 RCA|2626|Quarteto_em_C_menor|Beethov
396 en|Julliard|#WAR|23699|Touchstone|Co
432 rea|Corea|#
```

Índices

Denso × esparso

Denso

Um índice é chamado denso quando o número de entradas no índice e o número de registros de dados é o mesmo.
Neste caso, há uma correspondência 1 : 1 (uma entrada para cada registro).

Esparso

Um índice esparso é aquele que contém menor número de entradas que o número de registros de dados.
A relação é chamada 1 : n (uma entrada para n registros)

Índices

Exemplo: denso

Índice pela chave `artista`

Chave	Endereço
Bernstein	306
Corea	92
Ferras	348
Giulini	126
Julliard	43
Karajan	212
Leinsdorf	254
Maazel	0
Springsteen	168
Sweet Honey	396

Arquivo de dados (visão por registros)

```
000 LON|2312|Romeo_e_Julieta|Proko
030 fiev|Maazel|#RCA|2626|Quarteto
060 _em_C_menor|Beethoven|Julliard
090 |#WAR|23699|Touchstone|Corea|C
120 orea|#ANG|3795|Sinfonia_n._9|B
150 eethoven|Giulini|#COL|38358|Ne
180 braska|Springsteen|Springsteen
210 |#DG|18807|Sinfonia_n._9|Beeth
240 oven|Karajan|#MER|75016|Coq_d'
270 Or_Suite|Rimsky-Korsakov|Leins
300 dorf|#COL|31809|Sinfonia_n._9|
330 Dvorak|Bernstein|#DG|139201|Co
360 ncerto_de_Violino|Beethoven|Fe
390 rras|#FF|245|Good_News|Sweet_H
420 oney_Rock|Sweet_Honey|#
```

Índices

Exemplo: esparso

Índice (`gravadora+ID`)

Chave	Endereço
ANG3795	0
DG18807	176
RCA2626	360

Arquivo de dados

```
000 ANG|3795|Sinfonia_n._9|Beethoven|Giu
036 lini|#COL|31809|Sinfonia_n._9|Dvorak
072 |Bernstein|#COL|38358|Nebraska|Sprin
108 gsteen|Springsteen|#DG|139201|Concer
144 to_de_Violino|Beethoven|Ferras|#DG|1
180 8807|Sinfonia_n._9|Beethoven|Karajan
216 |#FF|245|Good_News|Sweet_Honey_Rock|
252 Sweet_Honey|#LON|2312|Romeo_e_Juliet
288 a|Prokofiev|Maazel|#MER|75016|Coq_d'
324 Or_Suite|Rimsky-Korsakov|Leinsdorf|#
360 RCA|2626|Quarteto_em_C_menor|Beethov
396 en|Julliard|#WAR|23699|Touchstone|Co
432 rea|Corea|#
```

Exercício

Discutam em grupos de dois a três alunos e respondam

- Como uma chave que não está no índice pode ser localizada?
- O que é necessário para que a busca funcione no caso de índice esparso?

Índices

Denso × esparso

Denso	Esparso
"Um para um"	"Um para muitos"
Armazenamento maior	Armazenamento menor
Agrupado ou não	Sempre agrupado
Ponteiro para o registro	Ponteiro para grupo de registros[1]
Possível "não-acesso" ao arquivo de dados	Acesso ao arquivo de dados necessário[2]

[1] Grupos são mapeados para blocos de disco

[2] No contexto de insucesso na busca

Índices

Índice primário

Em um índice primário, a ordenação da chave do arquivo de dados coincide com a ordenação das entradas de índice.

Índices primários

- Agrupados
- Densos ou esparsos

Índice primário não é um índice de chaves primárias

Índices
novo cenário – I

Novo cenário dos exemplos

Índices
novo cenário – II

Arquivo de dados: registros e campos de tamanhos fixos

```
000 ANG__3795____Beethoven__________
032 COL__31809___Dvorak_____________
064 COL__38358___Springsteen________
096 DG___139201__Beethoven__________
128 DG___18807___Beethoven__________
160 FF___245_____Sweet_Honey_Rock___
192 LON__2312____Prokofiev__________
224 MER__75016___Rimsky-Korsakov____
256 RCA__2626____Beethoven__________
288 WAR__23699___Corea______________
```

Campos: `gravadora` (4 bytes), `ID` (7 bytes), `artista` (19 bytes) – total: 32 bytes

Índices
novo cenário – III

Organização do disco rígido

- Blocos de 96 bytes

Considerações

- Três registros por bloco
- Unidade de transferência
  - 96 bytes por operação de entrada ou saída
  - Três registros por operação de entrada ou saída

Índices
novo cenário – IV

Arquivo de dados, separado por blocos

```
000 ANG__3795____Beethoven__________
032 COL__31809___Dvorak_____________
064 COL__38358___Springsteen________

096 DG___139201__Beethoven__________
128 DG___18807___Beethoven__________
160 FF___245_____Sweet_Honey_Rock___

192 LON__2312____Prokofiev__________
224 MER__75016___Rimsky-Korsakov____
256 RCA__2626____Beethoven__________

288 WAR__23699___Corea______________
320 $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
352 $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
```

Os cifrões ($) indicam espaço no bloco não usado pelo arquivo.

Índices

Exemplo: índice primário – agrupado e denso

Chave	Endereço
ANG3795	0
COL31809	32
COL38358	64
DG139201	96
DG18807	128
FF245	160
LON2312	192
MER75016	224
RCA2626	256
WAR23699	288

```
000 ANG__3795____Beethoven__________
032 COL__31809___Dvorak_____________
064 COL__38358___Springsteen________

096 DG___139201__Beethoven__________
128 DG___18807___Beethoven__________
160 FF___245_____Sweet_Honey_Rock___

192 LON__2312____Prokofiev__________
224 MER__75016___Rimsky-Korsakov____
256 RCA__2626____Beethoven__________

288 WAR__23699___Corea______________
320 $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
352 $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
```

Índices

Exemplo: índice primário – agrupado e esparso

Índice

Chave	End.	Bloco
ANG3795	0	0
DG139201	96	1
LON2312	192	2
WAR23699	288	3

```
000 ANG__3795____Beethoven__________
032 COL__31809___Dvorak_____________
064 COL__38358___Springsteen________

096 DG___139201__Beethoven__________
128 DG___18807___Beethoven__________
160 FF___245_____Sweet_Honey_Rock___

192 LON__2312____Prokofiev__________
224 MER__75016___Rimsky-Korsakov____
256 RCA__2626____Beethoven__________

288 WAR__23699___Corea______________
320 $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
352 $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
```

Exercícios I

Organizem-se em grupos de dois a quatro alunos

Considerem

- Registros de tamanho fixo com 128 bytes cada um
- Blocos de disco com tamanho de 4096 bytes
- Índice cujas entradas (chave e ponteiro) ocupem 16 bytes
- Um arquivo de dados com 31990 registros

Respondam

- Quantos registros cabem por bloco?
- Quantos blocos o arquivo de dados ocupa?
- Quantas entradas de índice cabem por bloco?

Exercícios II

Ainda no mesmo cenário, respondam

- Quantos blocos ocuparia um índice denso para este arquivo?
- Quantos blocos ocuparia um índice esparso para este arquivo?

Exercícios III

Considerem agora

- Registros de tamanho fixo com **100** bytes cada um
- Blocos de disco com tamanho de 4096 bytes

Respondam

- Quantos registros cabem por bloco?
- O que é feito com o espaço que "sobra"?

Índices

Índice secundário

O índice é dito secundário quando não há correspondência entre a ordem das chaves nos registros de dados e a ordem das entradas de índice.

Índices secundários

- Não agrupados
- Densos
- Múltiplos por arquivo de dados

Índices

Exemplo: índice secundário – não agrupado e denso

Chave	Endereço
ANG3795	128
COL31809	160
COL38358	288
DG139201	0
DG18807	224
FF245	64
LON2312	192
MER75016	32
RCA2626	96
WAR23699	256

```
000  DG___139201__Beethoven__________
032  MER__75016___Rimsky-Korsakov____
064  FF___245_____Sweet_Honey_Rock___
096  RCA__2626____Beethoven__________
128  ANG__3795____Beethoven__________
160  COL__31809___Dvorak_____________
192  LON__2312____Prokofiev__________
224  DG___18807___Beethoven__________
256  WAR__23699___Corea______________
288  COL__38358___Springsteen________
320  $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
352  $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
```

Exercício

Organizem-se em grupos de três a cinco alunos e discutam como funcionaria a **inserção de um novo registro** em um arquivo com índice, para os casos seguintes

- Índice primário denso
- Índice primário esparso
- Índice secundário

Créditos

Créditos

Jander Moreira
http://www.dc.ufscar.br/~jander

jander@dc.ufscar.br

Universidade Federal de São Carlos

http://www.ufscar.br

Departamento de Computação

http://www.dc.ufscar.br