

Organização e Recuperação da Informação

Indexação – Parte 2

Jander Moreira

UFSCar – DC

14 de setembro de 2017

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

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

Operações com
indexação

Arquivos sem
ordenação
Arquivos
ordenados

Créditos

Agenda

1 Índices

Múltiplos níveis
Repetição de chaves

2 Blocos

Registros de tamanho
fixo
Registros de tamanho
variável

Registros maiores que
blocos

3 Operações com indexação

Arquivos sem
ordenação
Arquivos ordenados

Índices

Múltiplos níveis
Repetição de
chaves

Blocos


Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos


Operações com
indexação

Arquivos sem
ordenação
Arquivos
ordenados

Créditos

Material de referência

 FOLK, M.; ZOELLICK, B. *File structures*. USA: Addison-Wesley Publishing Company, Inc., 1992.

 GARCIA-MOLINA, H.; ULLMAN, J. D.; WIDOM, J. *Implementação de sistemas de banco de dados*. New Jersey: Editora Campus, 2001.

 MOREIRA, J. *Armazenamento e recuperação da informação*. São Carlos: Coleção UAB–UFScar, 2011.

Jander Moreira

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

Operações com indexação

Arquivos sem
ordenação
Arquivos
ordenados

Créditos

Índices

Índices

- Múltiplos níveis
- Repetição de chaves

Blocos

- Registros de tamanho fixo
- Registros de tamanho variável
- Registros maiores que blocos

Operações com indexação

- Arquivos sem ordenação
- Arquivos ordenados

Créditos

Material desta seção inspirado em
Folk e Zoellick (1992), Capítulo 6

Índices

Múltiplos níveisRepetição de
chaves

Blocos

Registros de
tamanho fixoRegistros de
tamanho variávelRegistros maiores
que blocosOperações com
indexaçãoArquivos sem
ordenaçãoArquivos
ordenados

Créditos

Índice primário

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

- Agrupados
- Densos ou esparsos

Um índice primário é um arquivo ordenado pela chave de indexação

Índices

Múltiplos níveis

Repetição de
chaves

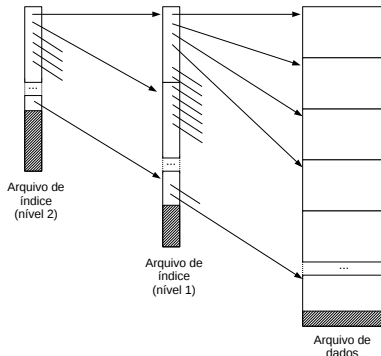
Blocos

Registros de
tamanho fixoRegistros de
tamanho variávelRegistros maiores
que blocosOperações com
indexaçãoArquivos sem
ordenaçãoArquivos
ordenados

Créditos

Índice primário em múltiplos níveis

- Nível de índice: índice para o índice
- Organização como índice esparsa ou densa



Índices

Índices

Múltiplos níveis

Repetição de
chaves

Blocos

Registros de
tamanho fixo

Registros de
tamanho variável

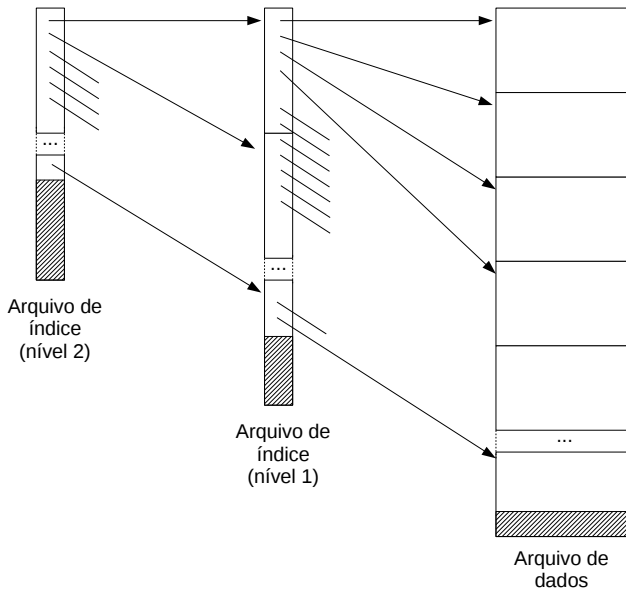
Registros maiores
que blocos

Operações com
indexação

Arquivos sem
ordenação

Arquivos
ordenados

Créditos



Exercício

Índices

Múltiplos níveis

Repetição de
chaves

Blocos

Registros de
tamanho fixo

Registros de
tamanho variável

Registros maiores
que blocos

Operações com indexação

Arquivos sem
ordenação

Arquivos
ordenados

Créditos

Considere

- Registros de dados com 230 bytes
- Entradas de índice com 20 bytes
- Blocos de 4096 bytes
- Dois níveis de índices

Com um único bloco no índice de nível 2, qual o número máximo de registros de que podem existir no arquivo de dados?

Índices

Múltiplos níveis

Repetição de
chaves

Blocos

Registros de
tamanho fixoRegistros de
tamanho variávelRegistros maiores
que blocosOperações com
indexaçãoArquivos sem
ordenaçãoArquivos
ordenados

Créditos

Índices

- Por chaves primárias
Sem repetições

Índice pela chave gravadora+ID

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

Índices

- Por chaves secundárias
Com provável repetição

Índice pela chave compositor

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

Índices

Múltiplos níveis

**Repetição de
chaves**

Blocos

Registros de
tamanho fixoRegistros de
tamanho variávelRegistros maiores
que blocosOperações com
indexaçãoArquivos sem
ordenaçãoArquivos
ordenados

Créditos

Soluções para repetições de chave no índice

- Índice denso com replicação das ocorrências
- Índice com apenas uma entrada por chave, mantendo-se os registros de dados em posições consecutivas
- Índice para lista de chaves primárias (“lista invertida”)

Índices

Múltiplos níveis

**Repetição de
chaves**

Blocos

Registros de
tamanho fixoRegistros de
tamanho variávelRegistros maiores
que blocosOperações com
indexaçãoArquivos sem
ordenaçãoArquivos
ordenados

Créditos

Índice com replicação de chaves

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

Índices

Múltiplos níveis

Repetição de
chaves

Blocos

Registros de
tamanho fixoRegistros de
tamanho variávelRegistros maiores
que blocosOperações com
indexaçãoArquivos sem
ordenaçãoArquivos
ordenados

Créditos

Índice com uma entrada por chave

Chave	Endereço
Beethoven	0
Corea	181
Dvorak	215
Prokofiev	257
Rimsky-Korsakov	300
Springsteen	352
Sweet Honey Rock	396

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

Necessidade de ordenação ou aglutinação dos dados

Índices

Múltiplos níveis

**Repetição de
chaves**

Blocos

Registros de
tamanho fixoRegistros de
tamanho variávelRegistros maiores
que blocosOperações com
indexaçãoArquivos sem
ordenaçãoArquivos
ordenados

Créditos

Listas invertidas

- Uso de chaves primárias
- Entradas de índice com chaves repetidas usam lista das chaves primárias



Índices

Múltiplos níveis

Repetição de
chaves

Blocos

Registros de
tamanho fixoRegistros de
tamanho variávelRegistros maiores
que blocosOperações com
indexaçãoArquivos sem
ordenaçãoArquivos
ordenados

Créditos

Arquivo de dados para listas invertidas

```

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 | #

```

Índices II

Índices

Múltiplos níveis

Repetição de
chaves

Blocos

Registros de
tamanho fixoRegistros de
tamanho variávelRegistros maiores
que blocosOperações com
indexaçãoArquivos sem
ordenaçãoArquivos
ordenados

Créditos

Lista de chaves primárias e endereços de registros para listas invertidas

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

Chave primária: gravadora+ID

Índices

Múltiplos níveis

Repetição de
chaves

Blocos

Registros de
tamanho fixoRegistros de
tamanho variávelRegistros maiores
que blocosOperações com
indexaçãoArquivos sem
ordenaçãoArquivos
ordenados

Créditos

Grupos de chaves primárias por chave secundária

Chave secundária	Chave primária
Beethoven	ANG3795
	DG139201
	DG18807
	RCA2626
Corea	WAR23699
Dvorak	COL31809
Prokofiev	LON2312
Rimsky-Korsakov	MER75016
Springsteen	COL38358
Sweet Honey Rock	FF245

Índices IV

Índices

Múltiplos níveis

Repetição de
chaves

Blocos

Registros de
tamanho fixoRegistros de
tamanho variávelRegistros maiores
que blocosOperações com
indexaçãoArquivos sem
ordenaçãoArquivos
ordenados

Créditos

Lista invertida associando cada chave secundária à sua lista

Índice		Lista invertida		
Chave secundária	Início da lista	Pos	Ch. prim.	Próximo
Beethoven	5	0	COL31809	≡
Corea	9	1	FF245	≡
Dvorak	0	2	ANG3795	7
Prokofiev	3	3	LON2312	≡
Rimsky-Korsakov	6	4	COL38358	≡
Springsteen	4	5	DG18807	2
Sweet Honey Rock	1	6	MER75016	≡
		7	DG139201	8
		8	RCA2626	≡
		9	WAR23699	≡

Índices

- Múltiplos níveis
- Repetição de chaves

Blocos

- Registros de tamanho fixo
- Registros de tamanho variável
- Registros maiores que blocos

Operações com indexação

- Arquivos sem ordenação
- Arquivos ordenados

Créditos

Blocos

Índices

- Múltiplos níveis
- Repetição de chaves

Blocos

- Registros de tamanho fixo
- Registros de tamanho variável
- Registros maiores que blocos

Operações com indexação

- Arquivos sem ordenação
- Arquivos ordenados

Créditos

Material desta seção inspirado em Garcia-Molina, Ullman e Widom (2001), capítulo 3 (seções 3.3.2 e 3.4.4)

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

Operações com
indexação

Arquivos sem
ordenação
Arquivos
ordenados

Créditos

Blocos

Blocos

Os blocos são as unidades de transferência de dados de entrada e saída.

Cada operação de acesso a disco implica na movimentação de blocos inteiros.

- Transferências de E/S são múltiplas do tamanho do bloco
- O mapeamento dos blocos do arquivo para os blocos de disco é transparente (via sistema operacional)

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

Operações com
indexação

Arquivos sem
ordenação
Arquivos
ordenados

Créditos

Eficiência no acesso a dados em arquivos

- Consideração do acesso aos blocos
- Mapeamento dos registros nos blocos

Visão do arquivo como um conjunto de blocos

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

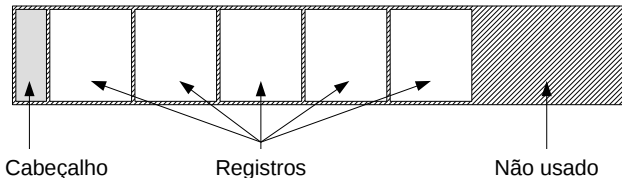
Operações com
indexação

Arquivos sem
ordenação
Arquivos
ordenados

Créditos

Uso do bloco de disco

- Uso de registros inteiros ou “quebrados” em blocos
- Controle interno do bloco (informações)
 - Número de registros
 - Deslocamentos dos registros
 - Ponteiros para outros blocos
 - Etc.



Índices

Múltiplos níveis
Repetição de
chaves

Blocos

**Registros de
tamanho fixo**

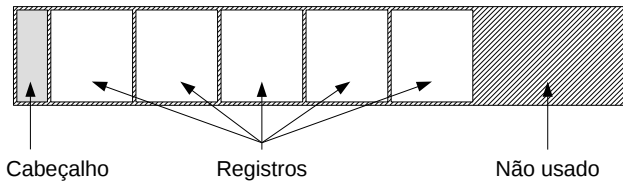
Registros de
tamanho variável
Registros maiores
que blocos

Operações com
indexação

Arquivos sem
ordenação
Arquivos
ordenados

Créditos

Blocos com registros de tamanho fixo



Índices

Múltiplos níveis
Repetição de
chaves

Blocos

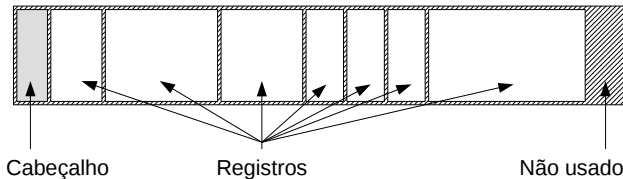
Registros de
tamanho fixo
**Registros de
tamanho variável**
Registros maiores
que blocos

Operações com
indexação

Arquivos sem
ordenação
Arquivos
ordenados

Créditos

Blocos com registros de tamanho variável



Índices

Múltiplos níveis
Repetição de
chaves

Blocos

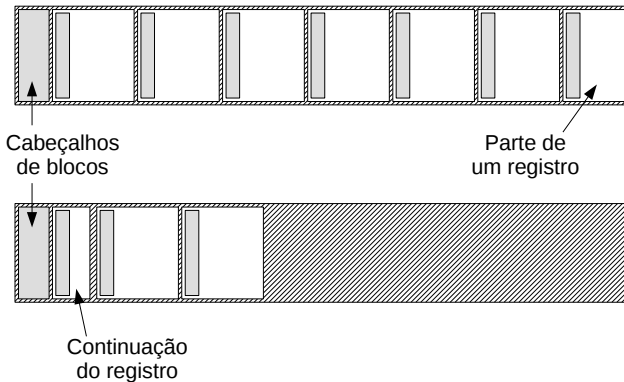
Registros de
tamanho fixo
Registros de
tamanho variável
**Registros maiores
que blocos**

Operações com
indexação

Arquivos sem
ordenação
Arquivos
ordenados

Créditos

Blocos com registros parcialmente armazenados



Índices

Múltiplos níveis
Repetição de
chaves

Blocos

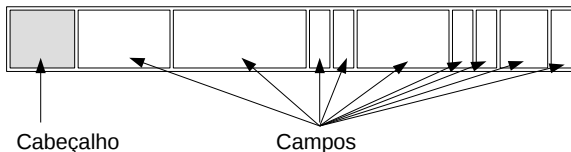
Registros de
tamanho fixo
Registros de
tamanho variável
**Registros maiores
que blocos**

Operações com
indexação

Arquivos sem
ordenação
Arquivos
ordenados

Créditos

Cabeçalho de registro



- Informações sobre os campos
- Informações sobre quebra do registro
- Ponteiros para os outros fragmentos do registro

Índices

- Múltiplos níveis
- Repetição de chaves

Blocos

- Registros de tamanho fixo
- Registros de tamanho variável
- Registros maiores que blocos**

Operações com indexação

- Arquivos sem ordenação
- Arquivos ordenados

Créditos

Exercício

Em um arquivo com registros de tamanho variável mantido em um arquivo ordenado por uma dada chave, é possível fazer pesquisa binária usando-se essa mesma chave?

Discuta a questão com seus colegas em pequenos grupos.

Índices

- Múltiplos níveis
- Repetição de chaves

Blocos

- Registros de tamanho fixo
- Registros de tamanho variável
- Registros maiores que blocos

Operações com indexação

- Arquivos sem ordenação
- Arquivos ordenados

Créditos

Operações com indexação

Operações com indexação

Índices

- Múltiplos níveis
- Repetição de chaves

Blocos

- Registros de tamanho fixo
- Registros de tamanho variável
- Registros maiores que blocos

Operações com indexação

- Arquivos sem ordenação
- Arquivos ordenados

Créditos

Material desta seção inspirado em Garcia-Molina, Ullman e Widom (2001), capítulo 3 (seção 3.5) e Moreira (2011), Unidade 5

Operações com indexação

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

Operações com indexação

Arquivos sem
ordenação
Arquivos
ordenados

Créditos

Operações de interesse

- Inserção
- Remoção
- Manutenção

Situações em que se aplicam

- Arquivos sem ordenação (dados)
- Arquivos ordenados (dados ou índice)

Contexto: arquivos com controle de blocos

Operações com indexação

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

Operações com indexação

Arquivos sem
ordenação
Arquivos
ordenados

Créditos

Operações em arquivos indexados

- Implicações de alterações no arquivo de dados nos arquivos de índice
- Cada operação no arquivo de dados implica no ajuste do índice

Operações com indexação

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

Operações com indexação

**Arquivos sem
ordenação**
Arquivos
ordenados

Créditos

Arquivos sem ordenação

Remoção

- ① Marcação como removido
 - Manutenção de lista de disponíveis
- ② Remoção física
 - Cópia do último registro do arquivo sobre o removido
 - Tratamento do último bloco (que cedeu o registro)

Operações com indexação

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

Operações com indexação

**Arquivos sem
ordenação**
Arquivos
ordenados

Créditos

Arquivos sem ordenação

Inserção

- 1 Uso da lista de posições disponíveis ou inserção no final
 - Tamanho fixo: uso do topo da lista
 - Tamanho variável: *{first,best,worst}-fit*
- 2 Inserção no final

Operações com indexação

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

Operações com indexação

**Arquivos sem
ordenação**
Arquivos
ordenados

Créditos

Arquivos sem ordenação

Manutenção (compactação)

- Varredura do arquivo de dados, com eliminação dos registros marcados como removidos
- Regeneração dos arquivos de índice

Operações com indexação

Índices

Múltiplos níveis

Repetição de
chaves

Blocos

Registros de
tamanho fixo

Registros de
tamanho variável

Registros maiores
que blocos

Operações com indexação

Arquivos sem
ordenação

**Arquivos
ordenados**

Créditos

Arquivos ordenados

Remoção

- 1 Marcação como removido
- 2 Remoção física
 - Deslocamento de registros (custo altíssimo)

Operações com indexação

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

Operações com indexação

Arquivos sem
ordenação
**Arquivos
ordenados**

Créditos

Arquivos ordenados

Inserção

- Determinação da posição (bloco) de inserção
 - Pesquisa binária
 - Tentativa de aproveitamento de espaço disponível no bloco
 - Alternativas se não houver espaço
 - 1 Deslocamento de registros (custo altíssimo)
 - 2 Uso de bloco de estouro (*overflow block*)
- Manutenção dos registros ordenados dentro do bloco

Bloco de estouro

- Bloco (em outro arquivo ou não)
- Existência de um ponteiro para o bloco de estouro no bloco original

Operações com indexação

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

Operações com indexação

Arquivos sem
ordenação
**Arquivos
ordenados**

Créditos

Arquivos ordenados

Manutenção (compactação)

- Varredura do arquivo de dados, com eliminação dos registros marcados como removidos
- Eliminação de espaços disponíveis dentro de blocos
 - ① Eliminação de blocos de estouro
 - ② Manutenção de espaço ocioso para novas inserções
- Regeneração dos arquivos de índice

Índices

- Múltiplos níveis
- Repetição de chaves

Blocos

- Registros de tamanho fixo
- Registros de tamanho variável
- Registros maiores que blocos

Operações com indexação

- Arquivos sem ordenação
- Arquivos ordenados

Créditos

Créditos

Índices

Múltiplos níveis
Repetição de
chaves

Blocos

Registros de
tamanho fixo
Registros de
tamanho variável
Registros maiores
que blocos

Operações com
indexação

Arquivos sem
ordenação
Arquivos
ordenados

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> >