

Índices

- Manter arquivos ordenados para permitir pesquisa binária → muito caro.
- Vantagens de se usar um índice simples para um arquivo de dados
 - adicionar registros muito mais rapidamente do que em um arquivo ordenado, desde que o índice seja suficientemente pequeno para ser mantido em memória principal.
- Vamos assumir que estas condições são válidas, e que o índice é carregado inteiro do disco (onde é mantido) para um vetor de registros denominado INDEX[].

1

Índices - exemplo

- índices simples** → consiste simplesmente de vetores com chaves e campos de referência.

Exemplo: Suponha que temos uma enorme coleção de CD's, e desejamos acessá-los através de um arquivo. Para cada registro, manter as seguintes informações:

Número: → identificação
Título
Artista
Rótulo → nome da gravadora
Ano

54	5
143	4
210	3
323	1
329	2
400	0

INDEX ↑
RRN

0	400 Minas Milton Nascimento Emi-Odeon 1975
1	323 Falso brilhante Elis Regina Philips 1976
2	329 A Arte de Chico Buarque PolyGram 1982
3	210 Chico Canta Chico Buarque Philips 1985
4	143 A Arte de Milton nascimento Universal 1988
5	54 Geraes Milton Nascimento Emi-Odeon 1976

2

Manutenção de índices

- Criar dois arquivos: de índice e de dados(em disco)**
Os arquivos de dados e de índices são criados inicialmente como arquivos vazios. Estes arquivos serão posteriormente carregados com seus dados.
- Carregar o arquivo de índice para memória no vetor INDEX[]**
→ leitura seqüencial do arquivo de índices
→ carregar em um vetor
- Buscar um registro**
→ fazer busca (binária) no vetor de índices
→ obter referência para arquivo de dados
→ ler registro no arquivo de dados
- Operações**
 - adição de registro**
 - a inserção deve ser feita no arquivo de dados e no índice.
 - o índice deve ser mantido ordenado → pode ser necessário reorganizá-lo.

3

Manutenção de índices

4. Operações (continuação)

eliminação de registro

- remove do arquivo de dados, usando algum dos mecanismos de remoção já vistos; e também do índice.
- A remoção do registro do índice pode exigir a sua reorganização, ou pode-se simplesmente marcar os registros como removidos, sem removê-los de fato.

atualização de registro: a atualização cai em duas categorias:

- Mudar o valor da chave:** pode exigir uma reorganização do arquivo índice e do arquivo de dados. Conceitualmente, pode ser tratada como uma remoção seguida de inserção de novo registro.
- Mudar o conteúdo:** se a chave não foi alterada → alterar o arquivo de dados, sem mexer no índice.

4

Manutenção de índices (cont)

3. Operações (continuação)

atualizar índice no disco

- Quando a cópia em memória foi alterada (inserções e/ou remoções).
- O que acontece se esta atualização não é feita, ou é feita apenas parcialmente em disco? (Ex: o programa não terminou adequadamente).

Deve haver um mecanismo que permita saber se o índice está atualizado.

Exemplo: Utilizando um flag "out-of-date"

- O flag é setado no arquivo índice mantido em disco assim que a sua cópia na memória é alterada.
- Esse flag pode ser mantido no registro header do arquivo índice, e deve ser setado assim que este é carregado na memória, e atualizado (resetado) sempre que o índice é reescrito no disco.
- Todo programa, antes de usar o índice, verifica o flag: se está setado, indica que o arquivo está desatualizado.
- se o programa detecta que o índice está desatualizado, deve existir uma função a ser ativada que reconstrua o índice a partir do arquivo de dados.

5

Índices muito grandes

Se o índice não cabe na memória, o seu acesso e manutenção precisa ser feito em memória secundária

- Não é mais aconselhável usar índices simples:
 - a busca binária pode exigir vários acessos (seek) a disco;
 - a necessidade de deslocar registros nas inserções e remoções de registros tornaria a manutenção do índice excessivamente cara.
- Utilizar:
 - uma organização em *hashing* para o índice (caso a velocidade de acesso seja a prioridade máxima); ou
 - árvores-B, caso se deseje combinar acesso por chaves e acesso seqüencial eficientemente.

6

Índice primário – Exemplo de custo de acesso

Tabela ordenada com $r = 30.000$ registros
 Disco com tamanho de bloco $B = 1024$ bytes
 Registros de tamanho fixo $R = 100$ bytes
 Fator de bloco $bfr = \lfloor B/R \rfloor = \lfloor 1024/100 \rfloor = 10$ registros por bloco
 Número de blocos do arquivo $b = \lceil r/bfr \rceil = \lceil 30.000/10 \rceil = 3.000$ blocos

Custo para pesquisa binária diretamente na tabela $\lceil \log_2 b \rceil = \lceil \log_2 3.000 \rceil = 12$ acessos

Supondo a construção de um índice principal para o arquivo acima

Tamanho campo chave ordenação $V = 9$ bytes
 Tamanho de um ponteiro de bloco $P = 6$ bytes
 Tamanho de cada entrada do índice $R_i = (9 + 6) = 15$ bytes
 Fator de bloco do índice $bfr_i = \lfloor B/R_i \rfloor = \lfloor 1024/15 \rfloor = 68$ entradas por bloco
 Número total de entradas do índice é igual ao número de blocos do arquivo de dados $r_i = 3.000$
 O número de blocos para o arquivo de índice é $b_i = \lceil r_i/bfr_i \rceil = \lceil 3.000/68 \rceil = 45$ blocos

Custo para pesquisa binária no arquivo índice $\lceil \log_2 b_i \rceil = \lceil \log_2 45 \rceil = 6$ acessos
 acesso a bloco adicional para a tabela de dados = 7 acessos

7

Índices para acesso por múltiplas chaves

- Buscas via chave primária são raras
 - Encontre o registro com ra 0218807!
- Rotineiramente se faz acesso a registros por chave secundária
 - Encontre os alunos que nasceram entre 1987 e 1999.
- Um índice secundário é um arquivo ordenado contendo 2 campos:
 - a chave de busca (campo de indexação) e
 - a chave primária.
- A chave de busca, neste caso, não ordena o arquivo de dados.
- O índice secundário referencia o índice primário e não o registro físico no arquivo de dados.

8

Índices para acesso por múltiplas chaves

- Pode-se criar tantos índices quantas sejam as chaves de acesso. É possível implementar vários índices secundários para um arquivo.
- Cada índice dá uma visão diferente de um arquivo de dados.
 - assim como um catálogo de biblioteca permite pesquisar livros por autor, título ou assunto.
- Utilizando esse tipo de índice pode-se realizar:
 - Buscas que sejam a **união** das respostas de índices secundários.
 - Buscas que sejam a **interseção** das respostas de índices secundários.

9

Índices secundários

Exemplo: Considerar o Index do exemplo anterior e criar índice para Artista

54	5	0	400 Minas Milton Nascimento Emi-Odeon 1975
143	4	1	323 Falso brilhante Elis Regina Philips 1976
210	3	2	329 A Arte de Chico Buarque PolyGram 1982
323	1	3	210 Chico Canta Chico Buarque Philips 1985
329	2	4	143 A Arte de Milton nascimento Universal 1988
400	0	5	54 Geraes Milton Nascimento Emi-Odeon 1976

INDEX

Index para Artista	Chave primária	Índice secundário
CHICO BUARQUE	210	→ Campos chave na forma canônica → no formato que será usado na busca.
CHICO BUARQUE	329	Neste caso:
ELIS REGINA	323	→ letras maiúsculas
MILTON NASCIMEN	54	→ Tamanho max=15 caracteres
MILTON NASCIMEN	143	→ Para exibir o nome no formato normal deve-se buscar o artista no arquivo de dados.
MILTON NASCIMEN	400	

10

Índices secundários Alterações nas operações básicas

Adição de registro de dados

- inserir as entradas correspondentes nos índices primário e secundário.
 - Nenhum problema se os índices estão na RAM.
 - pode exigir que registros sejam deslocados, para criar uma posição de inserção, ou que um vetor de ponteiros para as estruturas seja re-arranjado.
- o campo **chave** no índice secundário → armazenado na forma canônica
 - o valor pode ser truncado porque o tamanho da chave deve ser mantido fixo
 - a forma canônica deve levar em consideração esta restrição de tamanho para que a busca no índice funcione corretamente.
- Observações
 - Diferença entre os índices primário e secundário → no secundário pode ocorrer duplicação de chaves.
 - Chaves duplicadas devem ser mantidas agrupadas e ordenadas mantendo a mesma ordem em que aparecem no INDEX (chave primária).

11

Índices secundários Alterações nas operações básicas

Eliminação de registro de dados

- Remoção do registro do arquivo de dados e de todos os índices
- Rearranjo dos registros remanescentes nos índices primário e secundários.
- Alternativa para reduzir o rearranjo:
 - Atualizar apenas o índice primário - não eliminar a entrada correspondente ao registro do índice secundário → a busca retorna um valor inválido.
 - Como o índice secundário referencia o índice primário (e não o registro físico no arquivo de dados), se for feita uma busca por um registro já removido essa condição será acusada na busca pela chave primária feita no índice primário, e a não remoção da entrada do índice secundário.

12

Índices secundários

Alterações nas operações básicas

- **Vantagem na redução do rearranjo:**
 - economia de tempo substancial quando vários índices secundários estão associados ao índice primário.
 - Se os índices estiverem sendo mantidos em disco, reduz o número de seek.
- **Custo na redução do rearranjo no índice secundário** → O espaço ocupado por registros inválidos. Para reduzir o custo:
 - Fazer "coletas de lixo" periódicas nos índices secundários.
 - Se o arquivo de dados é muito volátil, utilizar árvore-B para a estrutura de índice secundário → permite remoção sem que seja necessário re-arranjar muitos registros.

13

Índices secundários

Alterações nas operações básicas

Atualização de registro de dados

- Índices secundários não fazem referências diretas ao arquivo de dados → a atualização de registros afeta os índices secundários apenas se as chaves secundárias ou primárias forem alteradas.
- Existem 3 situações possíveis:
 - a atualização alterou uma chave secundária → reordenar o índice secundário para esta chave.
 - a atualização alterou a chave primária → reordenar o índice primário e corrigir os índices secundários (os campos de referência) → a atualização dos índices secundários não requer reorganização
 - alteração em outros campos → não afeta nenhum dos índices.

14

Busca usando múltiplas chaves

Aplicações das chaves secundárias

- uso de uma ou mais chaves para localizar conjuntos de registros do arquivo de dados, fazendo uma busca em vários índices e uma combinação (AND e/ou OR) dos resultados.
- Com o uso de índices secundários, este tipo de busca é simples e rápida, pois as operações lógicas são realizadas nos arquivos de índices, reduzindo assim os seeks.

15

Busca usando múltiplas chaves

Exemplo: arquivo de CDs, e dois arquivos secundários para ele, com chaves Artista e Rótulo.

Algumas consultas possíveis:

- Encontre CD com número= 54 (acesso por chave primária);
- Encontre CD's com Artista= "Chico Buarque" (chave secundária: Artista);
- Encontre todos os CD's com Rótulo= "Philips" (chave secundária: Rótulo).
- Encontre Artista="Milton Nascimento" e Rótulo="Emi-Odeon"
 - Busca nos índices secundários: por Artista e por Rótulo
 - Fazer um AND dos resultados (referência à chaves primárias) retornadas.
 - Buscar no índice primário os endereços dos registros resultantes no arquivo de dados.
 - Fazer seeks no arquivo de dados para buscar os registros.

16

Busca usando múltiplas chaves

Aplicações das chaves secundárias

→ Exemplo: arquivo de CDs, e dois arquivos secundários para ele, com chaves **Artista** e **Rótulo**

0	400	Minas	Milton Nascimento	Emi-Odeon	1975
1	323	Falso brilhante	Elis Regina	Philips	1976
2	329	A Arte de	Chico Buarque	PolyGram	1982
3	210	Chico Canta	Chico Buarque	Philips	1985
4	143	A Arte de	Milton nascimento	Universal	1988
5	54	Geraes	Milton Nascimento	Emi-Odeon	1976

54	5
143	4
210	3
323	1
329	2
400	0

CHICO BUARQUE	210
CHICO BUARQUE	329
ELIS REGINA	323
MILTON NASCIMENT	54
MILTON NASCIMENT	143
MILTON NASCIMENT	400

EMI-ODEON	54
EMI-ODEON	400
PHILIPS	210
PHILIPS	323
POLYGRAM	329
UNIVERSAL	143

Índice primário

Índices secundários

17

Busca usando múltiplas chaves

0	400	Minas	Milton Nascimento	Emi-Odeon	1975
1	323	Falso brilhante	Elis Regina	Philips	1976
2	329	A Arte de	Chico Buarque	PolyGram	1982
3	210	Chico Canta	Chico Buarque	Philips	1985
4	143	A Arte de	Milton nascimento	Universal	1988
5	54	Geraes	Milton Nascimento	Emi-Odeon	1976

54	5
143	4
210	3
323	1
329	2
400	0

Resultado da consulta:

1. encontre CD com número= 54 (acesso pela chave primária);

a) Buscar no índice primário o RRN correspondente ao 54

5

b) Buscar no arquivo de dados, o registro com RRN = 5

54 | Geraes | Milton Nascimento | Emi-Odeon | 1976

18

Busca usando múltiplas chaves

0	400	Minas	Milton Nascimento	Emi-Odeon	1975
1	323	Falso brilhante	Elis Regina	Philips	1976
2	329	A Arte de	Chico Buarque	PolyGram	1982
3	210	Chico Canta	Chico Buarque	Philips	1985
4	143	A Arte de	Milton nascimento	Universal	1988
5	54	Geraes	Milton Nascimento	Emi-Odeon	1976

54	5
143	4
210	3
323	1
329	2
400	0

Resultado da consulta:

2. encontre CD's com Artista= "Chico Buarque"

a) Buscar no índice secundário Artista= "Chico Buarque"

CHICO BUARQUE	210
CHICO BUARQUE	329

b) Buscar no índice primário o RRN correspondente 210 e o 329

3
2

c) Buscar no arquivo de dados, os registros com RRN 3 e 2

210	Chico Canta	Chico Buarque	Philips	1985
329	A Arte de	Chico Buarque	PolyGram	1982

19

Busca usando múltiplas chaves

0	400	Minas	Milton Nascimento	Emi-Odeon	1975
1	323	Falso brilhante	Elis Regina	Philips	1976
2	329	A Arte de	Chico Buarque	PolyGram	1982
3	210	Chico Canta	Chico Buarque	Philips	1985
4	143	A Arte de	Milton nascimento	Universal	1988
5	54	Geraes	Milton Nascimento	Emi-Odeon	1976

54	5
143	4
210	3
323	1
329	2
400	0

Resultado da consulta:

3. encontre todos os CD's com Rótulo= "Philips"

a) Buscar no índice secundário Rótulo=Philips

PHILIPS	210
PHILIPS	323

b) Buscar no índice primário o RRN correspondente 210 e o 329

3
1

c) Buscar os registros com RRN 3 e 1 no arquivo

210	Chico Canta	Chico Buarque	Philips	1985
323	Falso brilhante	Elis Regina	Philips	1976

20

Busca usando múltiplas chaves

0	400	Minas	Milton Nascimento	Emi-Odeon	1975
1	323	Falso brilhante	Elis Regina	Philips	1976
2	329	A Arte de	Chico Buarque	PolyGram	1982
3	210	Chico Canta	Chico Buarque	Philips	1985
4	143	A Arte de	Milton nascimento	Universal	1988
5	54	Geraes	Milton Nascimento	Emi-Odeon	1976

54	5
143	4
210	3
323	1
329	2
400	0

Resultado da consulta 4

Encontre CD's com Artista="Chico Buarque" e Rótulo="Philips"

a. encontre CD's com Artista= "Chico Buarque"

CHICO BUARQUE	210
CHICO BUARQUE	329

b. encontre todos os CD's com Rótulo= "Philips"

PHILIPS	210
POLYGRAM	329

c. Faça AND dos resultados (referência às chaves primárias)

210

d. Buscar registro com Artista="Chico Buarque" e Rótulo="Philips"

210	Chico Canta	Chico Buarque	Philips	1985
-----	-------------	---------------	---------	------

21

Melhoria de índices secundários

Problemas com as estruturas de índices comuns:

- a repetição das chaves secundárias, que resulta em arquivos maiores (e, portanto, com menores chances de caber na memória);
- a necessidade de reordenação dos índices cada vez que um novo registro é inserido no arquivo, mesmo que esse registro tenha um valor de chave secundária já existente no arquivo.

Solução 1: Associar um vetor de tamanho fixo a cada chave secundária.

- não é necessário reordenar o índice a cada inserção de registro.
- essa facilidade fica limitada a um número fixo de repetições.
- ocorre fragmentação interna enorme - que talvez não compense a eliminação das chaves duplicadas...

Exemplo:

CHICO BUARQUE	210	329			
ELIS REGINA	323				
MILTON NASCIMENTEN	54	143	400		

22

Melhoria de índices secundários: Listas invertidas

Solução 2: Ligar à lista de referências - Listas invertidas

- cada chave secundária é relacionada a uma lista encadeada das chaves primárias referenciadas
- redefinir o índice secundário de forma que ele seja composto por registros com 2 campos: um campo chave, e um campo com o RRN do primeiro registro na lista invertida.
- As referências às chaves primárias associadas a cada chave secundária são mantidas em um arquivo seqüencial separado, organizado segundo a entrada dos registros.
- O índice de chaves secundárias funciona da seguinte forma:
 - cada chave secundária leva a uma ou mais chaves primárias (daí o termo "invertida" → a chave secundária leva à lista de chaves primárias, a qual por sua vez leva aos registros...)

23

Melhoria de índices secundários: Listas invertidas

Solução 2: Ligar à lista de referências - Listas invertidas

- Exemplo: Considerando o arquivo de dados de Cds.

400	Minas	Milton Nascimento	Emi-Odeon	1975
323	Falso brilhante	Elis Regina	Philips	1976
329	A Arte de	Chico Buarque	PolyGram	1982
210	Chico Canta	Chico Buarque	Philips	1985
143	A Arte de	Milton nascimento	Universal	1988
54	Geraes	Milton Nascimento	Emi-Odeon	1976

54	5
143	4
210	3
323	1
329	2
400	0

Índice primário

CHICO BUARQUE	210
CHICO BUARQUE	329
ELIS REGINA	323
MILTON NASCIMENTEN	54
MILTON NASCIMENTEN	143
MILTON NASCIMENTEN	400

CHICO BUARQUE	3
ELIS REGINA	1
MILTON NASCIMENTEN	5

Índice secundário

0	400	-1
1	323	-1
2	329	-1
3	210	2
4	143	0
5	54	4

Índice secundário com lista invertida

24

Melhoria de índices secundários: Listas invertidas

- **Vantagens:**
 - o índice secundário só é alterado quando um registro com uma chave não existente é inserido, ou quando uma chave existente é alterada (nome trocado);
 - operações de eliminação, inserção ou alteração de registros já existentes implicam apenas na mudança do arquivo da lista invertida.
 - a ordenação do arquivo de índice secundário é mais rápida (menos registros)
 - o arquivo com a lista invertida nunca precisa ser ordenado, pois é sequencial
- **Problema:**
 - registros associados não estão adjacentes - podem ser necessários vários *seeks* para recuperar a lista. Esse problema só existe se não for possível manter a lista na memória.

25

Índices seletivos

- Uma das vantagens de índices secundários é que eles podem ser utilizados para "dividir" um arquivo em partes, de modo a fornecer visões distintas de um mesmo arquivo.
 - Por exemplo: CD's lançados depois de 1980.
 - Essa facilidade é útil quando o conteúdo de um arquivo pode ser dividido naturalmente e logicamente em categorias.

26

Associação do índice ao endereço físico do registro

- **Em índices primários** → a associação ocorre no momento em que o arquivo é criado.
 - fornece acesso direto e, portanto, mais rápido, a um registro, dada a sua chave.
- **Em índices secundários** → associadas a um endereço apenas no momento em que são de fato usadas
 - Isso é possível visto que os índices não se referem diretamente ao endereço físico dos registros, mas ao índice primário.

27

Associação do índice ao endereço físico do registro

- Vantagens da associação apenas no último momento (*late-binding*)
 - mínima quantidade de reorganização quando os registros são adicionados ou excluídos.
 - Alterações importantes são feitas em um lugar em vez de em muitos lugares.
- Desvantagens da associação apenas no último momento
 - Isso implica em um acesso mais lento, principalmente se os índices estiverem em disco.
 - pesquisa binária no índice secundário +
 - pesquisa binária no índice primário

28

Exercícios

1- Escreva um algoritmo para uma função de busca de um registro em arquivos de índice simples.

Solução:

Se for de chaves primárias usar pesquisa binária.

Senão

- fazer busca binária no índice secundário
- obter todas as chaves acima ou abaixo da encontrada que sejam iguais e
- selecionar a chave primária do registro procurado.

29

Exercícios

2- Por que, para a eliminação de um registro do arquivo de dados, é possível eliminar o registro apenas do índice primário, e não do secundário?

Solução:

Porque o índice secundário não aponta para um registro físico e sim para o índice primário, assim quando a chave primária não existir (embora a chave secundária exista), saberemos que o registro não existe.

30



Exercícios

3- Explique a importância do flag "out-of-date" no header de um arquivo de índices utilizado num ambiente de multiprogramação.

Solução:

Esse flag é importante para garantir a consistência do arquivo de dados com o arquivo de índices, com vários programas que os acessam, pois indica que o arquivo de índices está atualizado de acordo com o arquivo de dados. Por exemplo, se houver uma queda do sistema, o flag indicará inconsistência caso o arquivo de índices não tenha sido atualizado (antes da queda). Neste caso, o arquivo de índices deve ser reconstruído, para garantir a consistência dos índices dos programas.

31



Exercícios

4- O que é uma lista invertida de índices? Quais são suas vantagens?

Solução:

É uma lista de chaves primárias associado a uma chave secundária.

Suas vantagens são:

- *eliminar a repetição das chaves secundárias (uma chave secundária aponta para uma lista de chaves primárias). Isso permite economizar espaço.*
- *inserção ou eliminação de um registro → não é necessário alterar a tabela de índices secundários (movimentar os elementos ou reordená-los). Somente a lista invertida onde a chave primária será eliminada ou inserida, será atualizada.*

32