

SCC0215 – Organização de Arquivos Anderson Canale Garcia

Material adaptado de:

Thiago A. S. Pardo

Leandro C. Cintra

M.C.F. de Oliveira

Cristina D. Aguiar

Contexto

32	LON 2312 Romeo and Juliet Prokofiev		
. 77	RCA 2626 Quartet in C Sharp Minor		
132	WAR 23699 Touchstone Corea		
167	ANG 3795 Symphony No. 9 Beethoven		
211	COL 38358 Nebraska Springsteen		
256	DG 18807 Symphony No. 9 Beethoven		
300	MER 75016 Coq d´or Suite Rimsky		
353	COL 31809 Symphony No. 9 Dvorak		
396	DG 139201 Violin Concerto Beethoven		
442	FF 245 Good News Sweet Honey In The		



- Em geral, um índice fornece mecanismos para localizar informações
 - Índice de um livro ou catálogo de uma biblioteca
 - Facilitam muito o trabalho de busca!

- Em arquivos
 - Permite <u>localizar registros rapidamente</u>
 - Não é necessário ordenar arquivo de dados, nem quando novos registros são adicionados

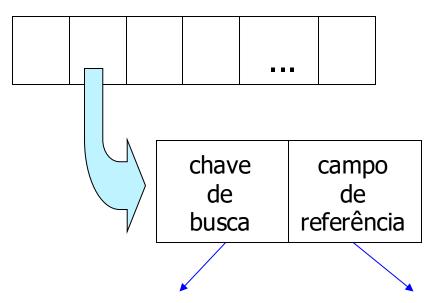


- Encontrar coisas
 - Registros em arquivos

 Impor uma ordem em um arquivo sem reorganizá-lo

Fornecer vários caminhos para um registro

Índice – Estrutura de dados



apesar de simples, índices proporcionam ferramentas poderosas para a recuperação de registros

valores ordenados RRN (registros de tamanho fixo)

byte offset (registros de tamanho variável)

Índice simples

- Exemplo: uma enorme coleção de CDs
- Registros de tamanho variável
 - ID: Número de identificação
 - Título
 - Compositor(es)
 - Artista
 - Rótulo: código da gravadora
- Chave primária: combinação de Rotulo e ID
 - Poderia ser qualquer outro campo ou combinação de campos que fosse único para cada registro

Endereço	Rótulo	ID	Título	Compositor	Artista
17	LON	2312	Romeo and Juliet	Prokofiev	Maazel
62	RCA	2623	Quartet in C Sharp Minor	Beethoven	Julliard
117	WAR	23699	Touchstone	Corea	Corea
152	ANG	3795	Symphony Nº 9	Beethoven	Giulini
196	COL	38358	Nebraska	Springsteen	Springsteen
241	DG	18807	Symphony No 9	Beethoven	Karajan
285	MER	75016	Cod d'Or Suite	Rimsky-Korsakov	Leinsdorf
338	COL	31809	Symphony No 9	Beethoven	Bernstein
382	DG	139201	Violin Concerto	Beethoven	Ferras
427	FF	245	Good News	Sweet Honey in the Rock	Sweet Honey in the Rock



- O índice consiste, em geral, em um outro arquivo com registros de tamanho fixo
 - Mesmo que o arquivo principal com os dados não tenha registros de tamanho fixo
- Cada registro do índice contém pelo menos 2 campos de tamanho fixo
 - Chave
 - Posição inicial (byte offset) ou RRN do registro no arquivo de dados

Exemplo

arquivo de índice

ANG3795	167
COL31809	353
COL38358	211
DG139201	396
DG18807	256
FF245	442
LON2312	32
MER75016	300
RCA2626	77
WAR23699	132

arquivo de dados

32	LON 2312 Romeo and Juliet Prokofiev		
77	RCA 2626 Quartet in C Sharp Minor		
132	WAR 23699 Touchstone Corea		
167	ANG 3795 Symphony No. 9 Beethoven		
211	COL 38358 Nebraska Springsteen		
256	DG 18807 Symphony No. 9 Beethoven		
300	MER 75016 Coq d´or Suite Rimsky		
353	COL 31809 Symphony No. 9 Dvorak		
396	DG 139201 Violin Concerto Beethoven		
442(FF 245 Good News Sweet Honey In The		

chave

Exemplo

arquivo de índice

ANG3795	167
COL31809	353
COL38358	211
DG139201	396
DG18807	256
FF245	442
LON2312	32
MER75016	300
RCA2626	77
WAR23699	132

arquivo de dados

32	LON 2312 Romeo and Juliet Prokofiev		
77	RCA 2626 Quartet in C Sharp Minor		•
132	WAR 23699 Touchstone Corea		
167	ANG 3795 Symphony No. 9 Beethoven		
211	COL 38358 Nebraska Springsteen		_
256	DG 18807 Symphony No. 9 Beethoven		
300	MER 75016 Coq d´or Suite Rimsky		
353	COL 31809 Symphony No. 9 Dvorak		
396	DG 139201 Violin Concerto Beethoven		
442	FF 245 Good News Sweet Honey In The		

valores ordenados

geralmente registros desordenados

Indice simples

- Vantagens do arquivo de índice sobre o de dados
 - Mais fácil de trabalhar, pois usa registros de tamanho fixo
 - Pode ser pesquisado com busca binária (em memória principal, inclusive, se valer a pena carregá-lo)
 - É muito menor do que o arquivo de dados
- Registros de tamanho fixo no arquivo índice impõem um limite ao tamanho da chave primária
 - O que acontece se a chave primária extrapolar o limite imposto?
- Os registros do índice poderiam conter outros campos além da chave/offset (por exemplo, o tamanho do registro)



- A inclusão de registros será muito mais rápida se o índice pode ser mantido em memória interna e o arquivo de dados é entry sequenced
- Dados a chave e o offset, um único seek é necessário no arquivo de dados para recuperar o registro correspondente

Índice

- Estrutura de dados
 - Chave de busca
 - Byte offset

Operações

Índice - Operações

- Criar o arquivo de índice vazio (criação)
- Carregar o índice em memória interna (carregamento)
- Pesquisa/busca
- Inserção
- Remoção
- Atualização
- Reescrita
- Destruição

Liste os dados de RCA2626

arquivo de índice

ANG3795	167
COL31809	353
COL38358	211
DG139201	396
DG18807	256
LON2312	32
MER75016	300
RCA2626	77
WAR23699	132

arquivo de dados

32	LON 2312 Romeo and Juliet Prokofiev	
77	RCA 2626 Quartet in C Sharp Minor	•
132	WAR 23699 Touchstone Corea	
167	ANG 3795 Symphony No. 9 Beethoven	
211	COL 38358 Nebraska Springsteen	
256	DG 18807 Symphony No. 9 Beethoven	
300	MER 75016 Coq d´or Suite Rimsky	_
353	COL 31809 Symphony No. 9 Dvorak	
396	DG 139201 Violin Concerto Beethoven	

Pesquisa/Busca

- Baseada na chave de busca
 - encontra a posição da chave no arquivo de índice
 - obtém o RRN ou o byte offset do registro correspondente à posição encontrada
 - encontra o registro no arquivo de dados
 - recupera o registro solicitado do arquivo de dados

Criação

- Cria o índice juntamente com a criação do arquivo de dados
 - apenas registro de cabeçalho
- Cria o índice baseado em um arquivo de dados já existente
 - registro de cabeçalho
 - demais registros (chave de busca + campo de referência), obtidos a partir de uma varredura no arquivo de dados

Inserção

 Adiciona registros no índice devido às inserções no arquivo de dados

inserção de um novo registro no arquivo de dados



inserção de um novo registro no arquivo de índice

arquivo não ordenado: realizada no final do arquivo ou com reaproveitamento de espaço necessidade de reorganização do índice, devido à ordenação da chave

Inserir

arquivo de índice

ANG3795	167
COL31809	353
COL38358	211
DG139201	396
DG18807	256
FF245	442
LON2312	32
MER75016	300
RCA2626	77
WAR23699	132

arquivo de dados

32	LON 2312 Romeo and Juliet Prokofiev		
77	RCA 2626 Quartet in C Sharp Minor		
132	WAR 23699 Touchstone Corea		
167	ANG 3795 Symphony No. 9 Beethoven		
211	COL 38358 Nebraska Springsteen		
256	DG 18807 Symphony No. 9 Beethoven		
300	MER 75016 Coq d´or Suite Rimsky		
353	COL 31809 Symphony No. 9 Dvorak		
396	DG 139201 Violin Concerto Beethoven		
442	FF 245 Good News Sweet Honey In The		

Remoção

 Remove registros no índice devido às remoções no arquivo de dados

remoção de um registro no arquivo de dados



remoção de um registro no arquivo de índice

 lógica: reaproveitamento de espaço lógica

 física: deslocamento dos registros

Atualização

- Modifica registros no índice devido às modificações no arquivo de dados
- Tratamento
 - remoção seguida de inserção
 ou
 - campo chave: reordenação do índice
 - campo não chave: ajuste do campo de referência se o registro mudar fisicamente no arquivo de dados

Destruição

- Exclui o arquivo de índice
- Demais funcionalidades relacionadas
 - realizadas diretamente sobre o arquivo de dados

Carregamento

- Carrega o arquivo de índice na memória principal antes de usá-lo
- Passos
 - aponta para o primeiro registro do arquivo de índice em disco
 - varre o arquivo de índices sequencialmente
 - cria o índice em memória principal, em geral implementado como um vetor

Reescrita

- Atualiza o arquivo de índice em disco com base no arquivo de índice em memória principal, quando necessário
- Informação adicional
 - status no registro de cabeçalho
 - verdadeiro/falso

inconsistência nos índices, devido à queda de energia, travamento do programa de atualização, etc.

Operações básicas no índice

- Para índices que cabem em memória
 - Criar arquivos índice e de dados
 - Carregar índice para memória
 - Inserir registro
 - Inserção deve ser feita no arquivo de dados...
 - e também no índice, que eventualmente será reorganizado
 - Eliminar registro
 - Remove do arquivo de dados
 - Remove também do índice
 - Índice pode ser reorganizado ou se pode apenas marcar os registros excluídos

Operações básicas no índice

- Para índices que cabem em memória
 - Atualizar registro duas categorias
 - Muda o valor da chave
 - Muda o conteúdo do registro
 - Atualizar índice no disco: caso sua cópia em memória tenha sido alterada
 - É imperativo que o programa se proteja contra índices desatualizados
 - Como?



- Deve haver um mecanismo que permita saber se o índice está atualizado em relação ao arquivo de dados
 - Possibilidade: um status 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 atualizado sempre que o índice é reescrito no disco
 - Se um programa detecta que o índice está desatualizado, uma função deve ser ativada para reconstruir o índice a partir do arquivo de dados



- Como saber qual é a chave primária do registro que se quer acessar?
- Normalmente, o acesso a registros não se faz por chave primária, e sim por chaves secundárias
 - Quando se procura a busca por um livro em um biblioteca, começa-se pelo seu número ou pelo título/autor?



Quais os dados do CD de código DG18807?



geralmente usado
internamente por um
programa, mas
raramente pelo usuário
de forma direta

Quais os dados da peça de Dvorak?



consulta típica de um usuário

Índice secundário

- Solução: cria-se um índice que relaciona uma chave secundária à chave primária (e não diretamente ao registro)
- Chave primária
 - Definido sobre um campo <u>sem</u> repetição
- Chave secundária
 - Definido sobre um campo com repetição

Índice Secundário

- Definido sobre uma chave secundária
 - podem ser definidos vários índices secundários para um mesmo arquivo de dados
- Tipos
 - fracamente ligado (Loosely Binding)
 - relaciona uma chave secundária à chave primária
 - fortemente ligado (Tight Binding)
 - relaciona uma chave secundária diretamente ao registro

Índice Secundário Fracamente Ligado

Beethoven	ANG3795
Beethoven	DG139201
Beethoven	DG18807
Beethoven	RCA2626
Corea	WAR23699
Dvorak	COL31809
Prokofiev	LON2312
Rimsky	MER75016
Springsteen	COL38358
Sweet Honey In The	FF245

ANG3795	167
COL31809	353
COL38358	211
DG139201	396
DG18807	256
FF245	442
LON2312	32
MER75016	300
RCA2626	77
WAR23699	132

32	LON 2312 Rom
77	RCA 2626 Quar
132	WAR 23699 Tou
167	ANG 3795 Sym
211	COL 38358 Nebr
256	DG 18807 Sym
300	MER 75016 Coq
353	COL 31809 Sym
396	DG 139201 Violin
442	FF 245 Good

arquivo de índice secundário

arquivo de índice primário

arquivo de dados

Busca

- Passo 1
 - pesquisar o índice de chave secundária para encontrar a chave primária relacionada
- Passo 2
 - usar a chave primária para pesquisar o índice de chave primária para encontrar o byte offset (ou RRN) do registro no arquivo de dados
- Passo 3
 - recuperar o registro desejado

Inserção

- Passo 1
 - inserir o registro no arquivo de dados
- Passo 2
 - inserir a entrada correspondente no arquivo de índice primário
- Passo 3
 - inserir a entrada correspondente em cada arquivo de índice secundário
 - chaves duplicadas devem ser mantidas agrupadas e ordenadas

Remoção

- Implica em remover o registro do arquivo de dados e de todos os índices
- Se índices mantidos ordenados, rearranjo dos registros remanescentes para não deixar "espaços vagos"

Remoção delete all references

- Passo 1
 - remover o registro no arquivo de dados
- Passo 2
 - remover a entrada correspondente no arquivo de índice primário
- Passo 3
 - remover a entrada correspondente em cada arquivo de índice secundário

Remoção

- Alternativa: atualizar apenas o índice primário, sem eliminar a entrada correspondente ao registro no índice secundário
- Vantagem: economia de tempo substancial quando vários índices secundários estão associados ao arquivo, principalmente se esses índices são mantidos em disco
- Custo: espaço ocupado por registros inválidos
 - Pode-se fazer "coletas de lixo" periódicas nos índices secundários
 - Ainda será um problema se o arquivo é muito volátil
 - Outra solução: índice em árvore-B

Remoção delete some references

- Passo 1
 - remover o registro no arquivo de dados
- Passo 2
 - remover a entrada correspondente no arquivo de índice primário

manutenção da entrada correspondente no arquivo de índice secundário

pesquisa no arquivo de índice primário acusará que o registro foi removido

Remoção

	delete all references	delete some references
vantagens	 sem queda de desempenho na busca por registros removidos índices permanecem do tamanho necessário 	 sem necessidade de reorganização a cada remoção economia de tempo nas remoções
desvantagens	 necessidade de reorganização a cada remoção processo altamente custoso, devido à ordenação 	 com queda de desempenho na busca na busca por registros removidos crescimento do tamanho dos índices e necessidade de reorganização periódica



- Atualizar registro 3 situações
 - Alterou uma chave secundária: o índice secundário para esta chave precisa ser reordenado
 - Alterou a chave primária: reordenar o índice primário e corrigir os campos de referência dos índices secundários
 - Vantagem: atualização dos índices secundários não requer reorganização
 - Alterou outros campos: não afeta nenhum dos índices
 - E se o tamanho do registro mudar?

Índice Secundário Fortemente Ligado

Beethoven	167
Beethoven	396
Beethoven	256
Beethoven	77
Corea	132
Dvorak	353
Prokofiev	32
Rimsky	300
Springsteen	211
Sweet Honey In The	442

32	LON 2312 Romeo and Juliet Prokofiev	
77	RCA 2626 Quartet in C Sharp Minor	
132	WAR 23699 Touchstone Corea	
167	ANG 3795 Symphony No. 9 Beethoven	
211	COL 38358 Nebraska Springsteen	
256	DG 18807 Symphony No. 9 Beethoven	
300	MER 75016 Coq d´or Suite Rimsky	
353	COL 31809 Symphony No. 9 Dvorak	
396	DG 139201 Violin Concerto Beethoven	
442	FF 245 Good News Sweet Honey In The	

Busca

- Passo 1
 - pesquisar o índice de chave secundária para encontrar o byte offset (ou RRN) do registro no arquivo de dados
- Passo 2
 - recuperar o registro desejado

Inserção

- Passo 1
 - inserir o registro no arquivo de dados
- Passo 2
 - inserir a entrada correspondente em cada arquivo de índice secundário
 - chaves duplicadas devem ser mantidas agrupadas e ordenadas

Remoção

- Passo 1
 - remover o registro no arquivo de dados
- Passo 2
 - delete all references: remover a entrada correspondente em cada arquivo de índice secundário

OU

 delete some references: manter a entrada correspondente em cada arquivo de índice secundário



- Alteração do valor da chave secundária
 - reordenação do índice secundário
- Alteração dos demais campos
 - não afeta nenhum dos índices

e se o tamanho do registro mudar?

Tipos de Índice Secundário

	fracamente ligado	fortemente ligado
vantagens	 diminui custo de remoções na abordagem delete some references modificação no arquivo de dados afeta apenas o índice primário menor complexidade de codificação 	 acesso direto índice primário → arquivo de dados índice secundário → arquivo de dados melhor desempenho na busca
desvantagens	 acesso indireto índice secundário → índice primário → arquivo de dados queda do desempenho na busca 	 alto custo para modificações modificação no arquivo de dados afeta todos os índices secundários maior complexidade de codificação



Acesso por múltiplas chaves

 Índices permitem muito mais do que simplesmente melhorar o tempo de busca por um registro

Múltiplos índices secundários

- Permitem manter diferentes visões dos registros em um arquivo de dados
- Permitem combinar chaves associadas e, deste modo, fazer buscas que combinam visões particulares



Busca usando múltiplas chaves

- Uma das <u>aplicações mais importantes das chaves</u> <u>secundárias</u> é localizar conjuntos de registros do arquivo de dados usando uma ou mais chaves
- Pode-se fazer uma busca em vários índices e combinar (AND,OR,NOT) os resultados
- Exemplo: encontre todos os registros de dados tal que
 - composer = "BEETHOVEN" AND title = "SYMPHONY NO. 9"

Busca usando múltiplas chaves

Indice por compositor

Índice por título

Chave secundária	Chave primária	Chave secundária	Chave primária
BEETHOVEN	ANG3795	COQ D'OR SUITE	MER75016
BEETHOVEN	DG139201	GOOD NEWS	FF245
BEETHOVEN	DG18807	NEBRASKA	COL31809
BEETHOVEN	RCA2626	QUARTET IN C SHARP	RCA2626
COREA	WAR23699	ROMEO AND JULIET	LON2312
DVORAK	COL31809	SYMPHONY NO 9	ANG3795
PROKOFIEV	LON2312	SYMPHONY NO 9	COL31809
RIMSKY-KORSAKOV	MER75016	SYMPHONY NO 9	DG18807
SPRINGSTEEN	COL38358	TOUCHSTONE	WAR23699
SWEET HONEY IN TH	FF245	VIOLIN CONCERTO	DG139201

Melhoria de índices secundários

- Dois problemas nas estruturas de índices vistas até agora
 - Repetição das chaves secundárias
 - Necessidade de reordenar os índices sempre 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
 - Limitado a um número fixo de repetições
 - Ocorre fragmentação interna enorme no índice que talvez não compense a eliminação da duplicação de chaves

Solução 1

 Associa um vetor de tamanho fixo a cada chave secundária

Beethoven	ANG3795	DG139201	DG18807	RCA2626
Corea	WAR23699			
Dvorak	COL31809			
Prokofiev	LON2312			
Rimsky	MER75016			
Springsteen	COL38358			
Sweet Honey In The	FF245			

Vetores de Tamanho Fixo

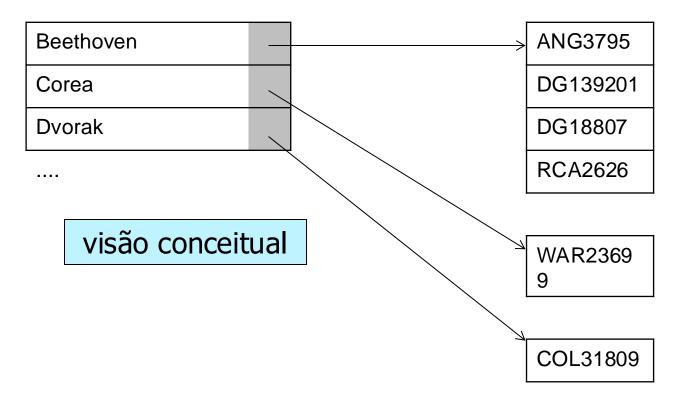
- Vantagem
 - não é necessário reordenar o índice secundário a cada inserção de chave secundária repetida
- Desvantagens
 - limitado a um número fixo de repetições
 - grande ocorrência de fragmentação interna no índice
 - pode não compensar a eliminação da duplicação de chaves



- Solução 2: manter uma lista de referências listas invertidas
 - Pode-se associar cada chave secundária a uma lista encadeada das chaves primárias referenciadas
 - Índice secundário passa a ser composto por registros com 2 campos: campo chave e campo com o RRN/byte offset do primeiro registro com essa chave na lista invertida
 - Referências às chaves primárias associadas a cada chave secundária são mantidas em um arquivo seqüencial separado, organizado segundo a entrada dos registros

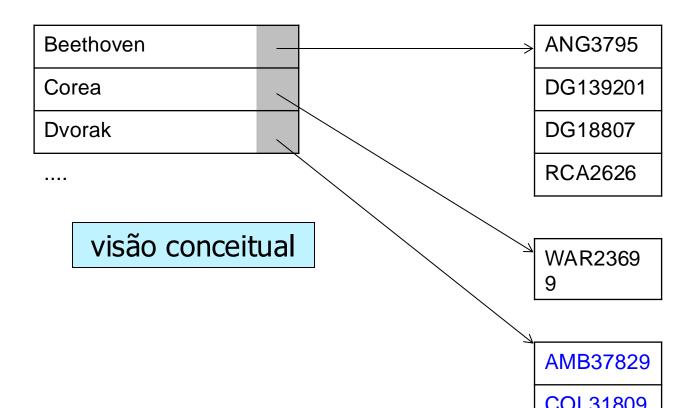
Solução 2

 Associa uma lista encadeada das chaves primárias a cada chave secundária



Listas Invertidas

 Inserção de um novo registro relativo a Dvorak



Listas Invertidas Implementação

Beethoven	3
Corea	2
Dvorak	10
Prokofiev	0
Rimsky	6
Springsteen	4
Sweet Honey In The	9

 Inserção de um novo registro relativo a Dvorak

0	LON2312	-1
1	RCA2626	-1
2	WAR23699	-1
3	ANG3795	8
4	COL38358	-1
5	DG18807	1
6	MER75016	-1
7	COL31809	-1
8	DG139201	5
9	FF245	-1
10	AMB37829	7

32	LON 2312 Rom
77	RCA 2626 Quar
132	WAR 23699 Tou
167	ANG 3795 Sym
211	COL 38358 Neb
256	DG 18807 Sym
300	MER 75016 Coq
	•••
353	COL 31809 Sym
396	DG 139201 Violi
442	FF 245 Good
530	AMB 37829 Bye
	•••

Listas invertidas

Vantagens desta estratégia?

Listas Invertidas Vantagens

- Indice secundário
 - alterado quando insere-se um registro com chave inexistente, ou quando altera-se chave já existente
- Remoção, inserção ou alteração de registros já existentes
 - alteração apenas no arquivo da lista invertida
 - modificação do campo de referência do índice se necessário
- Ordenação do arquivo de índice secundário
 - mais rápida: menos registros e registros menores
- Registros de tamanho fixo
 - facilita a adoção de um mecanismo para reaproveitamento de espaço

Listas invertidas

Problemas desta estratégia?

Listas Invertidas Desvantagem

- Chaves primárias associadas a uma certa chave secundária não estão adjacentes fisicamente no disco
 - pode ser necessário realizar vários seeks para recuperar a lista

Ideal: manter o índice e a lista na memória primária

Índices seletivos

- O índice não precisa cobrir todo o arquivo de dados
 - Índice de músicas clássicas
 - Índice de músicas lançadas depois de 1980
- Dependente da aplicação e uso dos dados

Índices muito grandes

- Se o índice não cabe inteiro na memória, o acesso e manutenção precisam ser feitos em memória secundária
- Não é mais aconselhável usar índices simples, uma vez que
 - A busca binária pode exigir vários acessos a disco
 - A necessidade de deslocar registros nas inserções e remoções de registros tornaria a manutenção do índice excessivamente cara

Índices muito grandes

- Utilizam-se <u>outras organizações</u>
 - Hashing, caso a velocidade de acesso seja a prioridade máxima
 - Acesso direto apenas
 - Árvores-B, caso se deseje combinar acesso por chaves e acesso seqüencial eficientemente



ATIVIDADES

Remova LON23112

arquivo de índice

ANG3795	167
COL31809	353
COL38358	211
DG139201	396
DG18807	256
FF245	442
LON2312	32
MER75016	300
RCA2626	77
WAR23699	132

arquivo de dados

32	LON 2312 Romeo and Juliet Prokofiev		
77	RCA 2626 Quartet in C Sharp Minor		
132	WAR 23699 Touchstone Corea		
167	ANG 3795 Symphony No. 9 Beethoven		
211	COL 38358 Nebraska Springsteen		
256	DG 18807 Symphony No. 9 Beethoven		
300	MER 75016 Coq d´or Suite Rimsky		
353	COL 31809 Symphony No. 9 Dvorak		
396	DG 139201 Violin Concerto Beethoven		
442	FF 245 Good News Sweet Honey In The		

Atualize LON para XXX

arquivo de índice

ANG3795	167
COL31809	353
COL38358	211
DG139201	396
DG18807	256
FF245	442
LON2312	32
MER75016	300
RCA2626	77
WAR23699	132

arquivo de dados

32	LON 2312 Romeo and Juliet Prokofiev	
77	RCA 2626 Quartet in C Sharp Minor	-
132	WAR 23699 Touchstone Corea	
167	ANG 3795 Symphony No. 9 Beethoven	
211	COL 38358 Nebraska Springsteen	
256	DG 18807 Symphony No. 9 Beethoven	
300	MER 75016 Coq d´or Suite Rimsky	_
353	COL 31809 Symphony No. 9 Dvorak	
396	DG 139201 Violin Concerto Beethoven	
442	FF 245 Good News Sweet Honey In The	