Algoritmos e Estruturas de Dados II

Profa. Debora Medeiros

Acesso a Registros

Adaptado dos Originais de:

Ricardo Campello

Thiago Pardo

Leandro C. Cintra

Maria Cristina F. de Oliveira

Acesso a registros

- Arquivos organizados por registros
  - Como buscar um registro específico?
    - Sem pensar em termos de primeiro, segundo...
    - Exemplo de nomes e endereços

Acesso a registros

- Arquivos organizados por registros
  - Como buscar um registro específico?
    - Sem pensar em termos de primeiro, segundo...
    - Exemplo de nomes e endereços
      - Chave

Chaves

- Uma **chave** (**key**) está associada a um registro e permite a sua recuperação
  - É uma ferramenta conceitual importante

Forma Canônica da Chave

- Quando se procura por "Ana"
  - "Ana", "ana" e "ANA" devem ser válidos para a busca
    - Converter chaves para uma forma canônica

Forma Canônica da Chave

- **Formas canônicas** para as chaves:
  - uma representação padronizada da chave
    - única, conforme com uma regra
  - por exemplo, todos os caracteres maiúsculos
    - Bastante utilizada
    - "Ana", "ana" e "ANA": forma canônica será **ANA**

Forma Canônica da Chave

- Ex: Quando se procura pela chave "Ana":
  - converter para "ANA" antes de realizar a busca

Chaves

- É desejável que sejam distintas
  - Identificar um único registro
- Senão...
  - Como resolver?
  - Exemplo dos endereços
    - Mostrar todos que se encaixam na busca?
    - Fornecer uma maneira de navegar pelos resultados?
  - Chave única atribuída quando os registros são adicionados: chave primária

Chaves Primária e Secundária

- Uma **chave primária** é, por definição, a chave utilizada para identificar unicamente um registro
  - Exemplos:
    - No. USP
    - CPF
    - RG
    - ...
  - Sobrenome, por outro lado, não é uma boa escolha ...

Escolha da Chave Primária

- A chave primária deve ser "dataless"
  - não deve ter um significado associado
  - não deve **mudar nunca**
    - Pode ser referenciada em relatórios, índices
  - Nome não é "dataless"
    - Apesar de (quase) nunca mudar

Escolha da Chave Primária

- Existência de significado poderia implicar mudança do valor da chave
  - invalidaria referências já existentes, baseadas na chave antiga
- Possibilidade de haver replicações
  - Caso do nome

Chaves Primária e Secundária

- Uma **chave secundária**, não necessariamente identifica unicamente um registro
  - pode ser utilizada para buscas por vários registros
    - por exemplo: todos os "**Silvas**" que moram em **São Paulo**
  - Útil para quando se quer recuperar registros por significado

Busca por um registro

- Como medir a performance?

Desempenho de Busca

- Na pesquisa em RAM, normalmente adotamos como medida do trabalho necessário o número de comparações efetuadas para obter o resultado da pesquisa
- Na pesquisa em arquivos, o acesso a disco é a operação mais cara e, portanto, o número de acessos a disco efetuados é adotado como medida do trabalho necessário para obter o resultado

Desempenho de Busca

- Na pesquisa em RAM, normalmente adotamos como medida do trabalho necessário o número de comparações efetuadas para obter o resultado da pesquisa
- Na pesquisa em arquivos, o acesso a disco é a operação mais cara e, portanto, o número de acessos a disco efetuados é adotado como medida do trabalho necessário para obter o resultado
  - Mecanismo de avaliação do custo associado ao método: contagem do número de chamadas à função de leitura de arquivo

Desempenho de Busca

- Assumimos (ingenuamente, por enquanto) que
  - Cada READ lê 1 registro e requer um seek
  - Todas as chamadas a READ tem o mesmo custo

Busca Sequencial

- Busca pelo registro que tem uma determinada chave, em um arquivo
  - Lê o arquivo, registro a registro, em busca de um registro contendo um certo valor de chave
  - Se cada registro lido demanda 1 acesso ao disco, tem-se custo de...

Busca Sequencial

- Busca pelo registro que tem uma determinada chave, em um arquivo
  - Lê o arquivo, registro a registro, em busca de um registro contendo um certo valor de chave
  - Se cada registro lido demanda 1 acesso ao disco, tem-se custo de **O(n) acessos**
    - n acessos para um arquivo com n registros

Busca Sequencial

- Maior custo de acesso ao disco?

Busca Sequencial

- Maior custo de acesso ao disco?
  - Seek
- Busca sequencial
  - seek-leitura → seek-leitura...
  - Como melhorar?

Busca Sequencial

- O custo de buscar e ler um registro, depois buscar e ler outro, é geralmente maior que o custo de buscar e ler dois registros sucessivos de uma só vez
  - apenas 1 *seeking* se ambos estiverem no mesmo cluster
- **Pode-se melhorar o desempenho da busca seqüencial lendo um bloco de registros por vez, e então processar este bloco em RAM**

Busca Sequencial

- Bloco
  - Mais um nível de hierarquia
    - Antes: registros, campos (organização lógica)
  - Porém, associado a desempenho
  - Relacionado com propriedades físicas do disco
    - Ex: tamanho do bloco múltiplo do tamanho do setor

Busca Sequencial

- Exemplo de Blocagem:
  - Arquivo com 4.000 registros
  - Busca sequencial por um registro, sem blocagem, requer, em média, **2.000** acessos
  - Se um bloco é capaz de armazenar, em média, 16 registros, o número médio de acessos cai para...

Busca Sequencial

- Exemplo de Blocagem:
  - Arquivo com 4.000 registros
  - Busca seqüencial por um registro, sem blocagem, requer, em média, **2.000** acessos
  - Se um bloco é capaz de armazenar, em média, 16 registros, o número médio de acessos cai para...
    - **125**
    - **Complexidade?**

Busca Sequencial

- Blocagem:
  - Cada acesso gasta um pouco mais de tempo, mas o ganho é considerável
    - redução do número de *seekings*
  - melhora o desempenho, mas o custo continua diretamente proporcional ao tamanho do arquivo, ou seja, **O(n) acessos**
    - somente constante é reduzida
  - Aumenta a quantidade de dados transferida
    - Mesmo quando o registro procurado é o primeiro, todo o bloco é transferido

Busca Sequencial

- Fácil de programar
- Requer estruturas de arquivos simples
- Aceitável ou Preferível em que casos?

Busca Sequencial

- Fácil de programar
- Requer estruturas de arquivos simples
- Aceitável ou Preferível:
  - Em arquivos com poucos registros
  - Em arquivos pouco pesquisados
  - Na busca por registros com um certo valor de chave secundária, para a qual se espera muitos registros (muitas ocorrências)
  - Na busca por todos os registros (e.g. mala direta)
- Felizmente, situações bastante frequentes

Acesso Direto

- A alternativa mais radical ao acesso sequencial é o **acesso direto** (ou **aleatório**)
- O acesso direto implica realizar um *seeking* lógico direto para o início do registro desejado
  - em geral, um único acesso ao disco traz o registro
- É **O(1) acessos** se:
  - a posição lógica do início do registro for conhecida
  - Como achar essa posição?

Acesso Direto

- Para localizar a posição exata do início do registro no arquivo lógico, pode-se utilizar um arq. de índice separado
  - associação entre chaves e posições dos registros
    - pode ser estruturado para otimizar consultas
    - veremos posteriormente no curso ...
- Ou um mapeamento funcional entre chaves e posições
  - tabela hash externa ...
- Ou pode-se ter uma chave **RRN**
  - ***RRN = Relative Record Number***
    - RRN = **0**, 1, 2, 3, ... (posição relativa do registro dentro do arquivo)

Acesso direto

- Como acessar diretamente com RRN?

Acesso Direto com RRN

- Acesso direto com **RRN** demanda o uso de registros de tamanho fixo T
  - Posição de início do registro (byte offset):
    - Byte Offset = RRN * T
  - Exemplo:
    - Registro na posição lógica 546 a partir do início
    - Tamanho de cada registro é 128
    - Byte offset = 546 x 128 = 69.888 bytes

Estruturas de registros

- Considerando tamanho fixo
  - Campos de tamanho fixo → OK
  - Ex:
    - soma = 30 bytes
    - Setor = 512
      - 32 bytes
  - Campos de tamanho variável

Estruturas de registros

- Campos de tamanho variável
  - Tamanho suficiente → simplicidade
  - Organização dentro do registro → utilização mais eficiente do espaço

Registro de tamanho fixo e campos de tamanho fixo:

Maria	Rua 1	123	São Carlos
João	Rua A	255	Rio Claro
Pedro	Rua 10	56	Rib. Preto

Registro de tamanho fixo e campos de tamanho variável:

Maria|Rua 1|123|São Carlos| ← Espaço vazio →|
João|Rua A|255|Rio Claro| ← Espaço vazio →|
Pedro|Rua 10|56|Rib. Preto| ← Espaço vazio →|

Estruturas de registros

- Em geral, é interessante manter algumas informações sobre o arquivo no seu início
  - cabeçalho no início do arquivo: **header record**

Registro Cabeçalho

- A existência de um registro cabeçalho torna um arquivo um objeto auto-descrito
  - O software pode acessar arquivos de forma mais flexível
  - Não precisa de informações prévias
  - Possibilita softwares de acesso a arquivo lidarem com uma maior variação de estruturas de arquivos
- Exemplos de informações armazenadas?

Registro Cabeçalho

- Algumas informações típicas são:
  - datas de criação e atualização
  - número de registros
  - tamanho de cada registro (caso fixo)
  - campos de cada registro (caso fixo)
    - no de campos
    - tipo de cada campo – inteiro, string com delimitador, etc
  - byte offsets de cada registro... índice "externo" !
- Bastante utilizado

Registro Cabeçalho

- Desvantagem dessa abordagem?
  - O software deve ser mais flexível e, portanto, sofisticado

Organização e Acesso de Arquivos

- **Organização de Arquivos**
  - campos de tamanho fixo ou variável
  - registros de tamanho fixo ou variável
  - técnicas de separação entre campos e registros
- **Acesso a arquivos**
  - acesso seqüencial
  - acesso direto

Organização e Acesso de Arquivos

- Considerações a respeito da organização:
  - arquivo pode ser dividido em campos ?
  - os campos podem ser agrupados em registros ?
  - campos e registros têm tamanho fixo ou variável ?
  - como separar os campos e registros ?
  - como identificar o espaço utilizado e o "lixo" ?
- Existem muitas respostas para estas questões...

Organização e Acesso de Arquivos

- Existem muitas respostas para estas questões ...
  - a escolha de uma organização em particular depende, entre outras coisas, do que se vai fazer com o arquivo
- Não necessariamente uma relação fixa:
  - Tamanho fixo → acesso direto
  - Tamanho variável → acesso sequencial

Organização e Acesso de Arquivos

- Por exemplo, arquivos com regs. de tamanhos muito diferentes deveriam utilizar regs. de tamanho variável
  - mas não é possível acessá-los diretamente por RRN
  - requer ao menos verificar um arquivo de índice ...
    - Complexidade de acesso por RRN será de O(1) acessos apenas se todo o arquivo de índice puder ser lido em memória de uma única vez

Organização e Acesso de Arquivos

- Pode ainda haver limitações de linguagem
- Exemplo:
  - C permite acesso a qualquer byte com **fseek**
    - permite implementar acesso direto a registros de tamanho variável uma vez que se conheça a posição de início do registro
  - Pascal permite *seeking* apenas para *records*
    - registros do mesmo tipo e tamanho
    - acesso direto a registros de tamanho variável não é viável

Observações

- Parte das informações atuais tratadas pelos computadores não se ajustam bem ao modelo de dados como seqüências de campos e registros
  - som, imagens, ...
- É mais fácil pensar em dados deste tipo como objetos que representam som, imagens, etc.
  - possuem sua própria maneira de serem manipulados

Observações

- O termo **modelo abstrato de dados** captura a noção de que o dado não precisa ser visto da forma como está armazenado e vice-versa
  - permite uma visão dos dados orientada à aplicação
- **Metadados** podem ser vistos como realizações desses modelos
  - são dados que descrevem os dados
    - arquivos com conteúdo auto-explicável (ex. PDF, PS, ...)
    - permite portabilidade e facilita conversões de padrões

Bibliografia

- **M. J. Folk and B. Zoellick, *File Structures: A Conceptual Toolkit*, Addison Wesley, 1987.**