Estrutras de Indexação de Arquivos

INF 220-Banco de Dados I – Prof. Jugurta Lisboa Filho

Apresentação elaborada por Glauber Costa (Estágio em Ensino 2011 – PPGCC/UFV)

Agenda

- Introdução
- Índices
- Tipos de índices
  - Índice primário
  - Índice de agrupamentos (clustering)
  - Índice secundários
- Índices multiníveis
- Árvores de pesquisa e árvores B*-tree
- Estruturas de arvores B*-tree

Objetivos

- Compreender a função dos índices em banco de dados, suas estruturas de armazenamento e formas de implementação.

Introdução

- Índices:
  - Estruturas de acesso auxiliares que são utilizadas para agilizar a recuperação de registros em resposta a certas condições de pesquisa.
  - Oferecem caminhos de acesso secundários a registros sem afetar o posicionamento físico do arquivo.
- Tipos de índices:
  - Baseados em arquivos ordenados (índices de único nível)
  - Estruturas de dados em árvores (índices multiníveis, B*-trees).
- Outras estruturas de construção:
  - Hashing ou outras estruturas de dados de pesquisa.

Índices

- Uma analogia:
  - Índices remissivos em livros:
    - Termos ordenados alfabeticamente
    - Associado a cada termo, temos uma indicação da página onde o termo ocorre
    - Vantagem:
      - Mais eficiente que folear todo o livro, buscando o termo.
  - Índices de banco de dados
    - Armazena o campo de índice junto com uma lista de ponteiros para todos os blocos de disco que contêm registros para esse valor de campo
    - Vantagem:
      - Como o arquivo de índices é menor e encontra-se ordenado, efetuar uma busca binária é bem mais eficiente que uma busca linear.

Tipos de índices

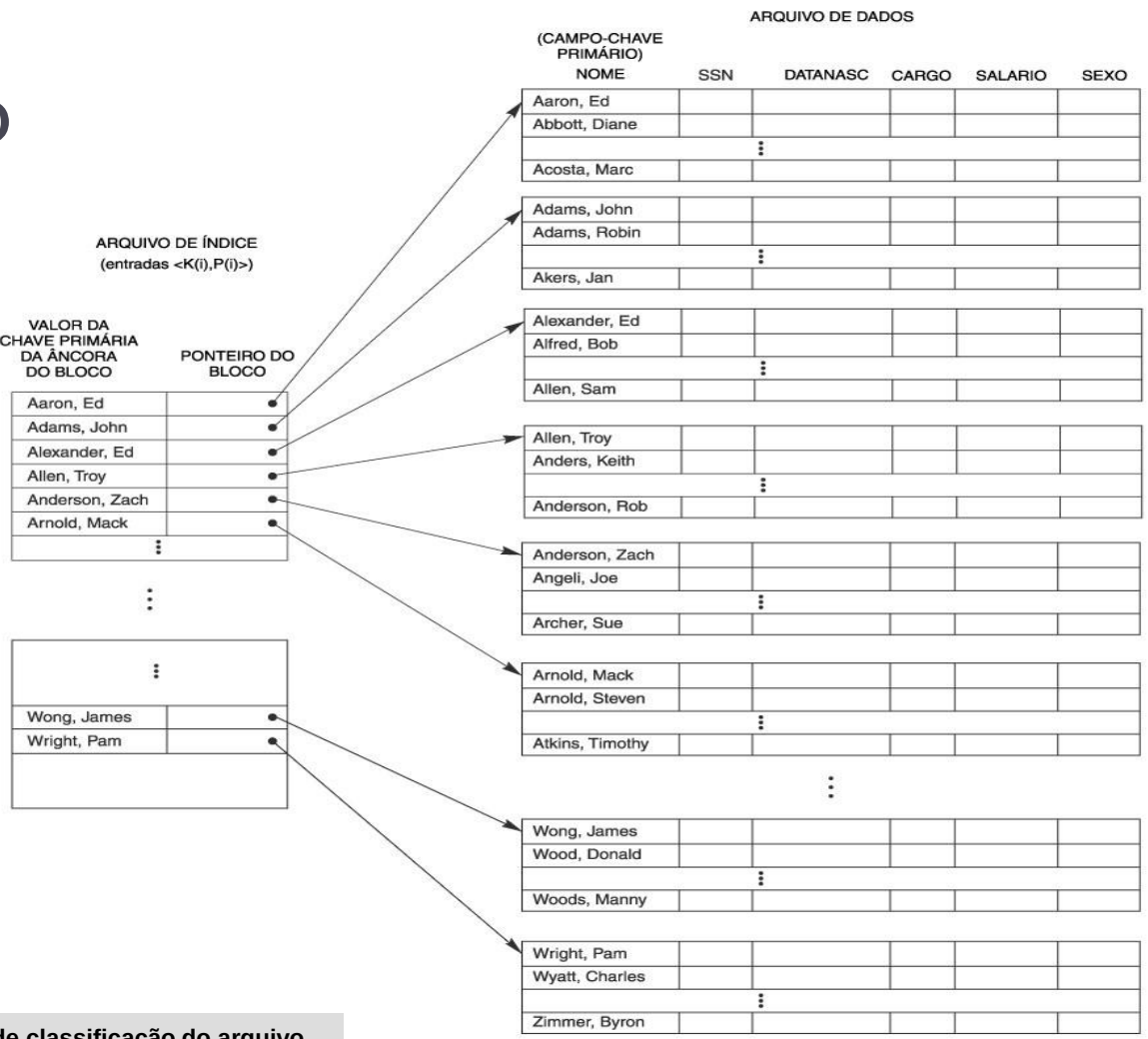
- Índice primário:
  - Especificado pelo campo chave da tabela
- Índice de agrupamento (clustering):
  - Se o campo de ordenação não for um campo chave
- Índice secundário:
  - Pode ser especificado em qualquer campo não ordenado de um arquivo.

Índice Primário

- Índice primário
  - É um arquivo ordenado com registros de tamanho fixo com 2 campos.
  - Existe registro de índice para cada bloco no arquivo de dados
- Estrutura
  - <K(i),P(I)>:
    - K = A chave de ordenação
    - P = Ponteiro para um bloco de disco:
      - O primeiro registro de cada bloco é chamado de registro de âncora ou âncora do bloco

Índice Primário

- Entradas:
  - <
    - K(1) = Aaron, Ed
    - P(1) = endereço do bloco 1
  - >
  - <
    - K(2) = Adams, John
    - P(2) = endereço do bloco 2
  - >
  - <
    - K(3) = Alexander, Ed
    - P(3) = endereço do bloco 3
  - >



Índice primário para o campo-chave de classificação do arquivo

Índice Primário

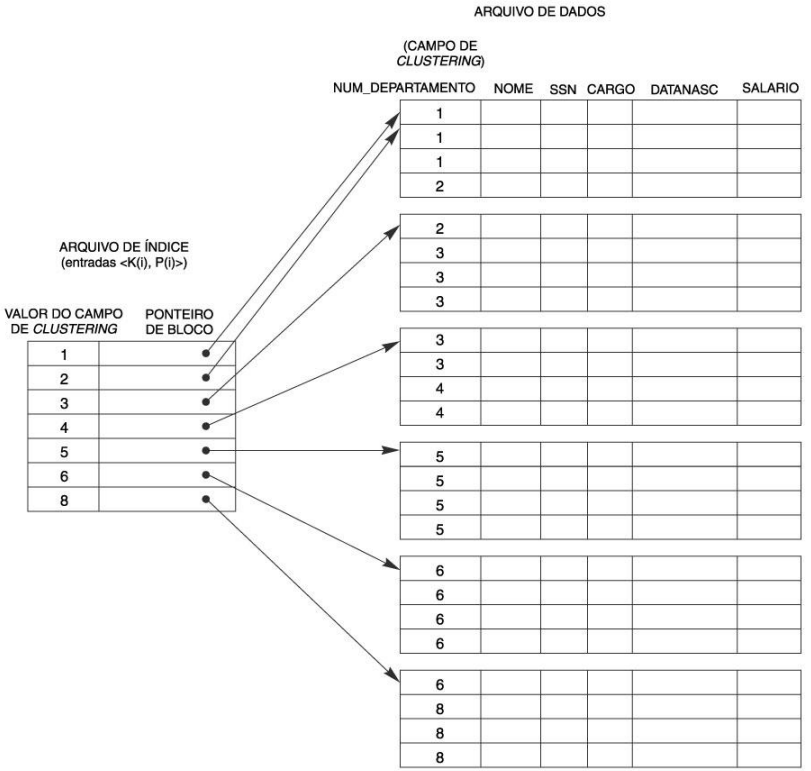
- Os índices se subdividem em:
  - Índices denso
    - Tem uma entrada de índice para cada chave de pesquisa;
    - Número de registros do arquivo de índice é igual ao número de registros do arquivo de dados;
  - Índices esparsos:
    - Tem entrada de índice somente para alguns valores de pesquisa;

Índice de Agrupamento (clustering)

- Arquivo agrupado:
  - Quando os registros de arquivo forem fisicamente ordenados em um campo não chave, chamado de **campo de agrupamento**.
  - Pode haver repetição dos **campos de agrupamento**
- Índice de agrupamento:
  - Índice não denso
  - Difere do índice primário pois não exige que o campo indexado tenha valor único

Índice de Agrupamento (clustering)

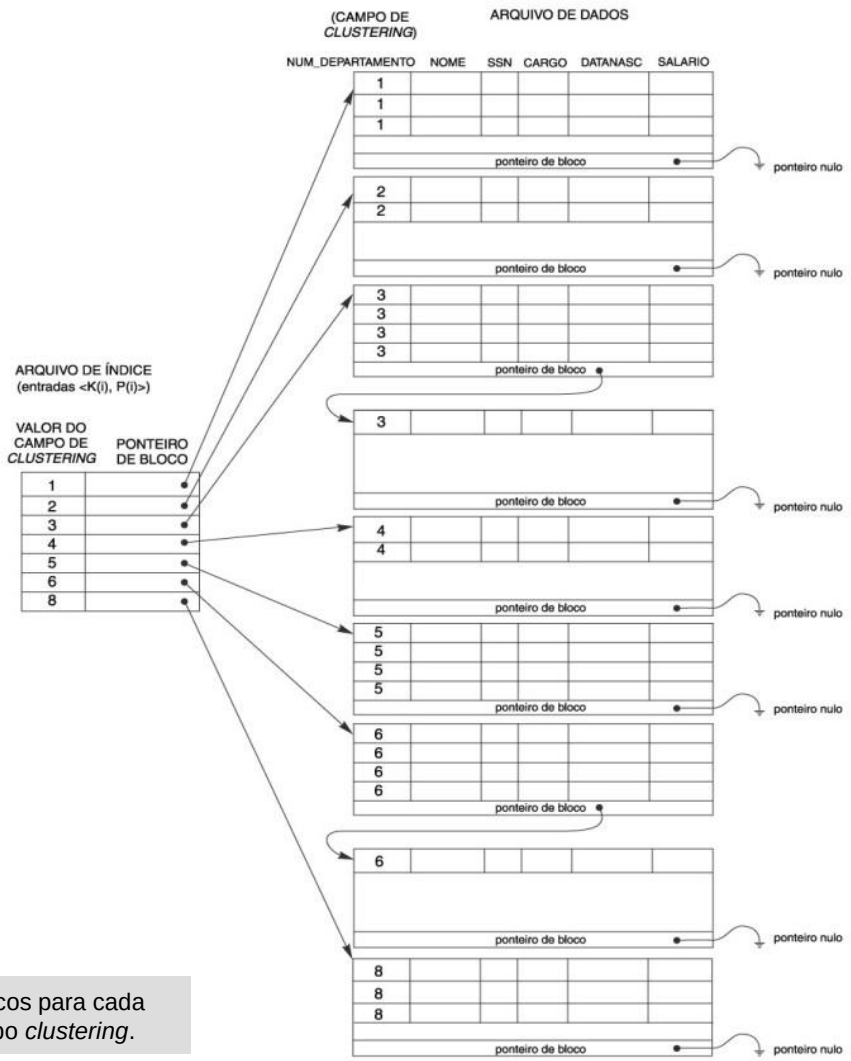
- Estrutura
  - <K(i),P(I)>:
    - K = A chave de ordenação (para cada valor distinto)
    - P = Ponteiro para um bloco de disco
- Problemas:
  - Inserção e exclusão causam problemas pois os registos de dados estão fisicamente ordenados
- Ideias diminuir o problema?



Um índice *clustering* para o campo NUM_DEPARTAMENTO**, que não é campo-chave de classificação, de um arquivo** EMPREGADO.

Índice de Agrupamento (clustering)

- Solução:
  - Reservar um bloco inteiro (ou cluster de blocos) para cada valor do campo de agrupamento.



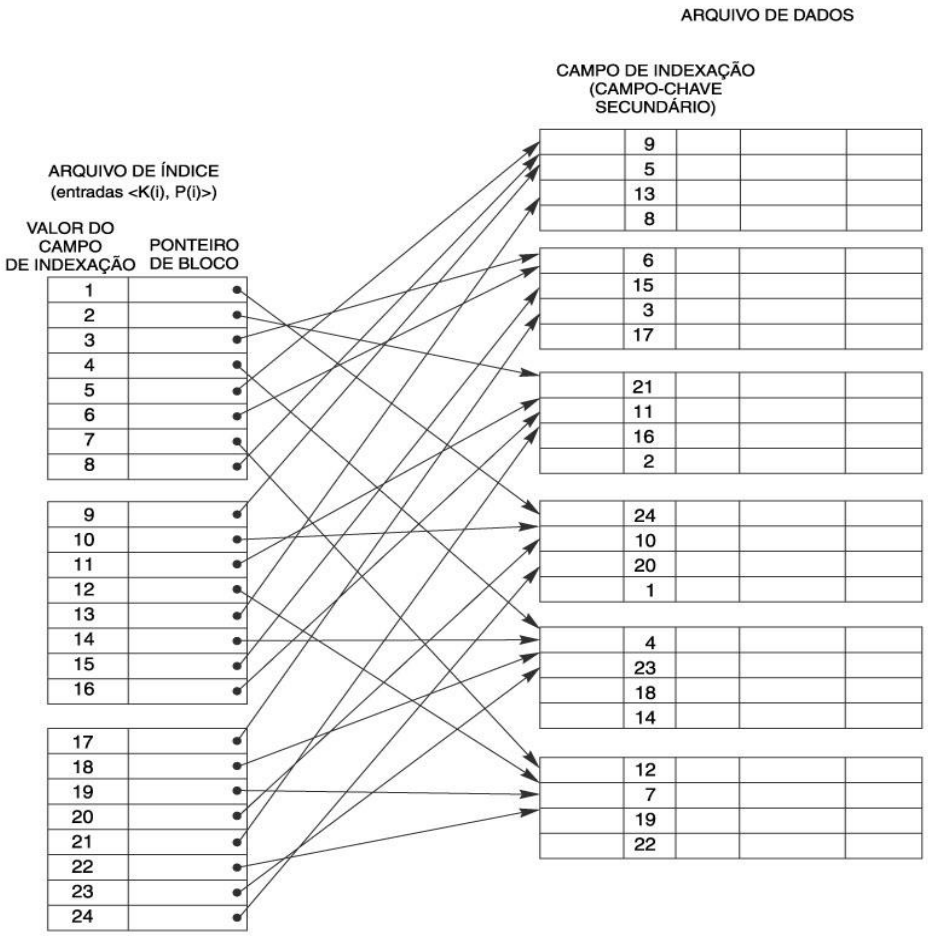
Um índice *clustering* com um grupo (*cluster*) separado de blocos para cada grupo de registros que compartilhem o mesmo valor de campo *clustering*.

Índice Secundário

- Índice secundário:
  - Meio secundário de acesso a dados de arquivos que já possuem um índice primário;
  - Os registros podem ser ordenados, desordenados ou *hashed*.
  - Os índices secundários podem ser criados:
    - A partir de uma chave candidata (com valor único)
    - A partir de um campo com chaves duplicadas.

Índice Secundário

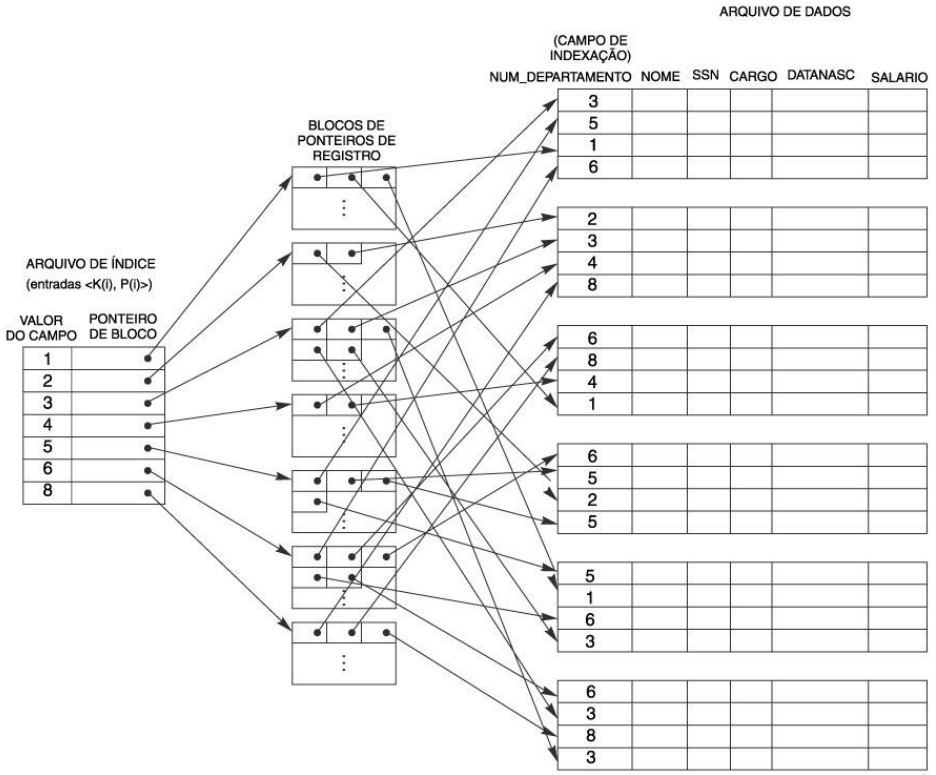
- Usando chaves candidatas
  - Segue a mesma ideia do índice primário:
    - É um arquivo ordenado com registros de tamanho fixo com 2 campos.
    - Existe registro de índice para cada bloco no arquivo de dados
  - Estrutura
    - <K(i),P(I)>
    - Como os registros não são ordenados fisicamente, não existem âncoras de bloco, logo é um índice denso



Um índice secundário denso (com ponteiros de bloco) em um campo não é chave de classificação de um arquivo.

Índice Secundário

- Usando chaves com repetição:
  - Diversos registros no arquivo de dados podem ter o mesmo valor
  - 3 opções de solução:
    1. Incluir entradas de índice duplicadas com o mesmo k(i).
    2. Ter registros de tamanho variável para entradas de índice, com um campo repetitivo para ponteiro. Mantendo uma lista de ponteiros: <P(i,1), P(i,2),...,P(i,k)>
    3. Manter as entradas do índice em tamanho fixo, com uma entrada para cada índice e criar um nível de indireção.



Um índice secundário(com ponteiros de registro), em um campo que não é campo-chave, implementado em um nível adicional, indireto, de forma que as entradas de índice sejam de tamanho fixo e possuam valores de campo únicos.

Índice Secundário

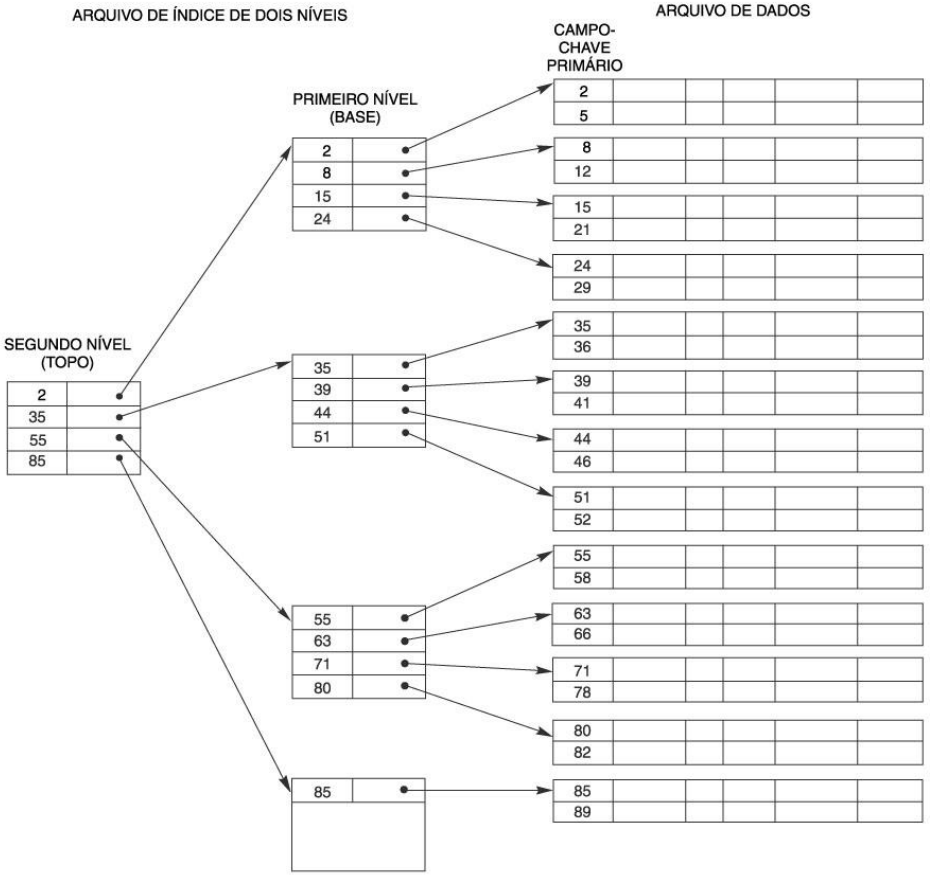
- Índice secundário x índice primário
  - Índices primários ocupam menos espaço
  - Índices secundários fornecem uma melhor relação custo x benefício pois teríamos que fazer uma busca linear no arquivo de dados se o índice secundário não fosse definido.

Índices Multiníveis

- Ordenação de um nível:
  - Aplica-se pesquisa binária ao índice para localizar ponteiros de blocos
  - Requer aproximadamente ($\log_2 b_i$) acessos de bloco
- Índices multiníveis:
  - Fator de bloco ou fator de blocagem *(bfr)* = *(B/R)*, onde:
    - B: tamanho do bloco em bytes;
    - R: tamanho do registro em bytes;
  - Permite reduzir a parte do índice que continuamos a pesquisar a partir do fator do bloco para o índice, que é maior que 2.
  - Reduz assim o espaço de busca muito mais rapidamente:
    - Pesquisa binária, a cada passo, reduz a metade o espaço
    - Com o índice mulitinível é possível dividir, a cada passo, em até n vezes o espaço de busca.

Índices Multiníveis

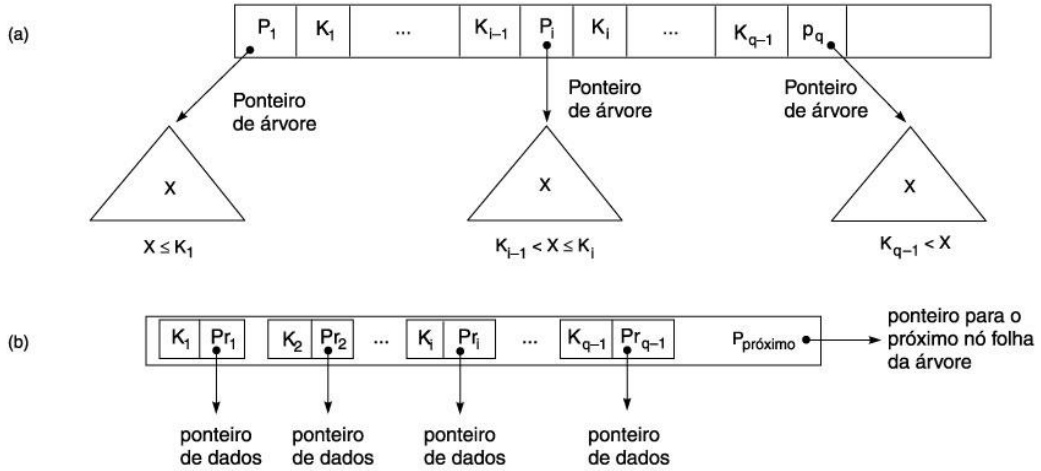
- Níveis de índice:
  - Primeiro nível:
    - Um arquivo ordenado com um valor distinto para cada K(i);
  - Segundo nível:
    - Como o primeiro nível é um arquivo ordenado, cria-se um índice primário a partir dele, podendo assim usar âncoras de bloco;
    - O fator de bloco para o segundo e para os demais níveis é o mesmo para os índices de primeiro nível.
- Só exige-se um segundo nível se o primeiro nível precisar de mais de um bloco de armazenamento.



Um índice primário de dois níveis que se parece com a organização ISAM (*Indexed Sequential Acess Method*).

Árvores de pesquisa e B*-trees

- Árvore de pesquisa:
  - Tipo especial de árvore utilizada para orientar a pesquisa por um registro, dado o valor de um dos campos do registro.
  - Os valores de um campo de índice nos guiam para o próximo nó até que alcancemos algum bloco do arquivo de dados com os registros solicitados.
- B-tree (ou árvore-B):
  - Possui restrições adicionais que garantem que a árvore esteja sempre balanceada.
- B+tree (ou árvore-B+):
  - Em uma B+tree os ponteiros de dados são armazenados apenas nos nós folha, logo a estrutura dos nós folha difere dos nós internos.



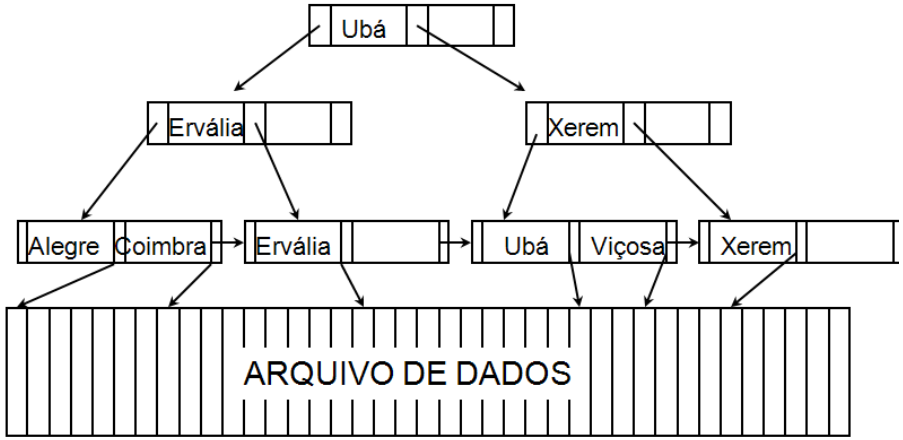
Os nós de uma árvore-B+.

(a) Nó interno de uma árvore-B+ com q – 1 valores de busca.

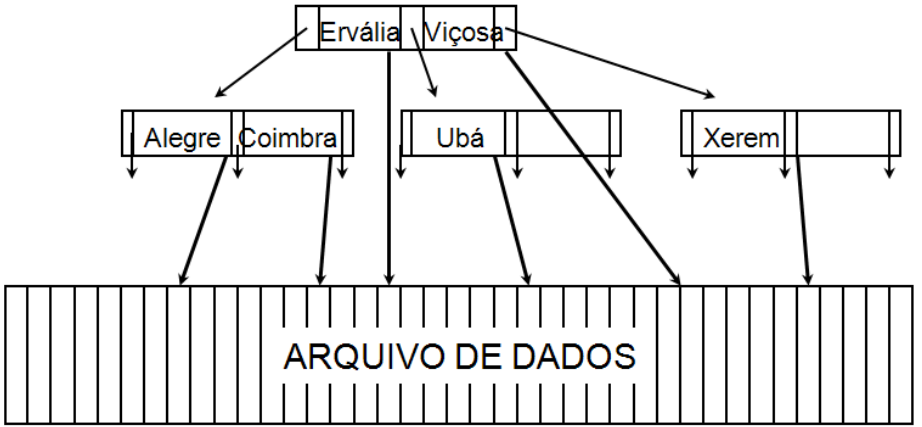
(b) (b) Nó folha de uma árvore-B+ com q - 1 valores de busca e q - 1 ponteiros de dados.

Estruturas de uma B*-tree

- Estrutura de uma árvore B+



- Estrutura de uma árvore B



Bibliografia

- Elmasri, R.; Navathe, S. – Sistemas de Banco de Dados – 6ª. Edição. São Paulo - 2011
- Lisboa Filho, J. – Notas de aula – Universidade Federal de Viçosa.