

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação
Disciplina de Organização de Arquivos
Profa. Dra. Cristina Dutra de Aguiar Ciferri
PAE Rayner de Melo Pires (Turma A)
PAE Anderson Chaves Carniel (Turma B)

Trabalho Prático

Este trabalho tem como objetivo realizar o armazenamento de registros em arquivos e realizar a recuperação eficiente desses registros usando um índice árvore-B*.

O trabalho deve ser feito em grupo de 4 alunos. A solução deve ser proposta exclusivamente pelo grupo com base nos conhecimentos adquiridos ao longo das aulas. Consulte as notas de aula e o livro texto quando necessário.

Primeira Parte do Trabalho (Parte I): peso 40%

Descrição do arquivo de dados. Cada registro do arquivo de dados deve conter dados relacionados a:

Turma A. FILME:

- idFilme (i.e. número único que identifica cada filme): campo numérico
- tituloFilme (i.e. nome do filme): campo alfanumérico
- descFilme (i.e. descrição do filme): campo alfanumérico
- producao (i.e. nome do país que produziu o filme): campo alfanumérico
- anoLancamento (i.e. ano de lançamento do filme): campo numérico
- duracaoFilme (i.e. duração do filme): campo numérico
- generoFilme (i.e. gênero do filme, como comédia, comédia romântica, drama): campo alfanumérico

Turma B. SÉRIE:

- idSérie (i.e. número único que identifica cada série): campo numérico
- tituloSérie (i.e. nome da série): campo alfanumérico
- descSérie (i.e. descrição da série): campo alfanumérico
- producao (i.e. nome do país que produziu a série): campo alfanumérico
- anoLancamento (i.e. ano de lançamento do primeiro episódio da série): campo numérico
- temporada (i.e. número de temporadas da série): campo numérico
- generoSérie (i.e. gênero da série, como comédia, comédia romântica, drama): campo alfanumérico

Organização do arquivo de dados. Deve ser considerada a organização híbrida de campos e registros, da seguinte forma:

- Campos de tamanho fixo e campos de tamanho variável com delimitadores entre campos.
- Registros de tamanho variável com delimitadores entre registros.

Programa. Implemente um programa que ofereça as seguintes funcionalidades.

[1] Ofereça uma interface por meio da qual o usuário possa escolher a funcionalidade a ser realizada.

[2] Realize a geração automática de 100 registros do arquivo de dados. O arquivo de dados não pode estar ordenado, nem pelo campo que é chave primária.

[3] Permita a recuperação dos dados, de todos os registros, armazenados no arquivo de dados, mostrando os dados de forma organizada na saída padrão para permitir a distinção dos campos e registros. Deve-se permitir a visualização dos registros, um por vez.

[4] Permita a recuperação dos dados de um registro com base em um valor do campo que é chave primária por meio da busca sequencial.

Restrições da Parte I. As seguintes restrições têm que ser garantidas no desenvolvimento da Parte I do trabalho.

[1] O arquivo de dados deve ser gravado em disco no **modo binário**. O modo texto não deve ser usado.

[2] Os dados do registro descrevem os nomes dos campos, os quais não podem ser alterados. Ademais, todos os campos devem estar presentes na implementação, e nenhum campo adicional pode ser incluído. O tamanho de cada campo deve ser determinado pelo grupo. Essa escolha deve ser explicada na **documentação externa**.

[3] O campo que é chave primária não deve armazenar valores repetidos de dados.

[4] Com exceção do campo que é chave primária, todos os demais campos devem representar campos de chave secundária, ou seja, os campos podem armazenar valores repetidos de dados.

[5] Os dados gerados devem ser os mais próximos da realidade. O grupo pode procurar por geradores de dados na internet ou então percorrer páginas de sítios de filmes/séries para a obtenção de dados de interesse desses sítios.

Segunda Parte do Trabalho (Parte II) : peso 60%

Descrição do índice. O índice árvore-B* é uma variação do índice árvore-B que implementa a **redistribuição durante a inserção**.

Definição formal. O índice árvore-B* com ordem m é definido formalmente como descrito a seguir.

1. Cada página (ou nó) do índice árvore-B* deve ser, pelo menos, da seguinte forma:
 $\langle P_1, \langle C_1, P_{R1} \rangle, P_2, \langle C_2, P_{R2} \rangle, \dots, P_{q-1}, \langle C_{q-1}, P_{Rq-1} \rangle, P_q \rangle$, onde ($q \leq m$) e
 - Cada C_i ($1 \leq i \leq q-1$) é uma chave de busca.
 - Cada P_{Ri} ($1 \leq i \leq q-1$) é um campo de referência para o registro no arquivo de dados que contém o registro de dados correspondente a C_i .
 - Cada P_j ($1 \leq j \leq q$) é um ponteiro para uma subárvore ou assume o valor nulo caso não exista subárvore.
2. Dentro de cada página
 - $C_1 < C_2 < \dots < C_{q-1}$.
3. Para todos os valores X da chave na subárvore apontada por P_i :
 - $C_{i-1} < X < C_i$ para $1 < i < q$
 - $X < K_i$ para $i = 1$
 - $K_{i-1} < X$ para $i = q$.
4. Cada página possui um máximo de m descendentes.
5. Cada página, exceto a raiz e as folhas, possui no mínimo $(2m-1)/3$ descendentes (*taxa de ocupação*).
6. A raiz possui pelo menos 2 descendentes, a menos que seja um nó folha.
7. Todas as folhas aparecem no mesmo nível.
8. Uma página não folha com k descendentes possui $k-1$ chaves.
9. Uma página folha possui no mínimo $\lfloor (2m-1)/3 \rfloor$ chaves e no máximo $m - 1$ chaves (*taxa de ocupação*).

Organização do arquivo de índice. Deve ser considerada a seguinte organização:

- Campos de tamanho fixo.
- Registros de tamanho fixo.

Programa. Estenda o programa da Parte I para que ele ofereça as funcionalidades descritas a seguir.

[1] Toda a vez que um registro for inserido no arquivo de dados, a sua entrada correspondente seja inserida no arquivo de índice árvore-B*. Ou seja, ocorre a inserção registro a registro no arquivo de dados, e elemento a elemento no arquivo de índice.

[2] Permita a recuperação de um registro do arquivo de dados com base em um valor do campo que é chave primária por meio do uso do índice árvore-B*.

[3] Permita a visualização gráfica do índice árvore-B*.

Restrições da Parte II. As seguintes restrições têm que ser garantidas no desenvolvimento da Parte II do trabalho.

[1] O arquivo de índice deve ser gravado em disco no **modo binário**. O modo texto não deve ser usado.

[2] O índice árvore-B* deve ser definido sobre o campo que é chave primária do arquivo de dados.

[3] Cada página (ou nó) do índice árvore-B* deve ser, pelo menos, da forma descrita no item **Definição Formal.1**. Adicionalmente, o grupo pode armazenar outros campos, de acordo com as necessidades de sua implementação. Os nomes de todos os campos bem como seus tamanhos devem ser determinados pelo grupo. Essa escolha deve ser explicada na **documentação externa**.

Restrições Globais

As seguintes restrições têm que ser garantidas no desenvolvimento do trabalho, tanto na Parte I quanto na Parte II.

[1] Os integrantes do grupo devem constar como comentário no início do código (i.e. NUSP e nome de cada integrante do grupo). Não será atribuída nota ao aluno cujos dados não constarem no código fonte.

[2] Todo código fonte deve ser documentado. A **documentação interna** inclui, dentre outros, a documentação de procedimentos, de funções, de variáveis, de partes do código fonte que realizam tarefas específicas. Ou seja, o código fonte deve ser documentado tanto em nível de rotinas quanto em nível de variáveis e blocos funcionais.

[3] A interface pode ser feita em modo texto (terminal) ou modo gráfico e deve ser funcional.

[4] A implementação deve ser realizada usando a linguagem de programação C. As funções das bibliotecas `<stdio.h>` devem ser utilizadas para operações relacionadas à escrita e leitura dos arquivos. A implementação não deve ser feita em qualquer outra linguagem de programação. O programa deverá compilar no GCC versão 4.8.2 ou superior.

[5] O programa deve ser acompanhado de uma **documentação externa**. A documentação externa deve conter a descrição dos principais conceitos usados no trabalho prático, as decisões de projeto e as suas justificativas (ex.: estruturas de dados e algoritmos usados), assim como qualquer outra consideração adicional assumida no desenvolvimento do trabalho prático. Em detalhes, a documentação externa deve possuir:

- **CAPA**, com as seguintes informações: o nome da instituição, o nome do curso, o nome da disciplina, o nome do professor responsável, o nome do trabalho prático, o nome dos participantes e os respectivos números USP, e a data de entrega do trabalho prático.
- **ÍNDICE**, listando os nomes das seções que compõem o trabalho prático e as suas respectivas páginas de início.
- **SEÇÕES 1 a N**: Quaisquer decisões de projeto. Em detalhes, a documentação referente a essas seções deve conter a descrição dos principais conceitos usados no trabalho prático, incluindo desenhos que facilitem a compreensão das estruturas de dados, as decisões de projeto e as suas justificativas, assim como qualquer outra consideração adicional assumida no desenvolvimento do trabalho prático. Todas as funcionalidades do programa devem ser descritas em detalhes. Por exemplo, na Parte I do trabalho, deve ser feita a descrição dos campos dos registros do arquivo de dados, contendo os nomes dos campos, os tamanhos dos campos e um desenho que mostra visualmente a estrutura dos registros. A escolha dos tamanhos dos campos deve ser justificada. Na Parte II do trabalho, deve ser feita a descrição das páginas do índice árvore-B*, contendo os nomes dos campos, os tamanhos dos campos e um desenho que mostra visualmente a estrutura da página. A escolha dos tamanhos dos campos, bem como dos campos propriamente ditos, deve ser justificada. Coloque aqui também os sistemas operacionais que foram usados e como o programa deve ser compilado e executado.
- **SEÇÃO N+1**: Cópias de telas da interface, por meio das quais é possível entender o funcionamento do programa.
- **SEÇÃO N+2**: Baterias de testes, as quais devem gerar informações que permitam acompanhar a execução do programa.
- **REFERÊNCIAS BIBLIOGRÁFICAS**, caso necessário.

Fundamentação Teórica

Conceitos e características dos diversos métodos para representar os conceitos de campo e de registro em um arquivo de dados podem ser encontrados nas transparências de sala de aula e também nas páginas 96 a 107 do livro *File Structures (second edition)*, de Michael J. Folk e Bill Zoellick.

Conceitos e características do índice árvore-B* podem ser encontrados nas transparências de sala de aula e também nas páginas 333 a 404 do livro *File Structures (second edition)*, de Michael J. Folk e Bill Zoellick.

Material para Entregar

Arquivo compactado. Deve ser preparado um arquivo .zip contendo:

- Código fonte do programa devidamente documentado.
- Makefile para a compilação do programa.
- Bibliotecas necessárias para a execução do programa.
- Documentação externa em formato .pdf.

Instruções de entrega.

[1] Enviar o arquivo compactado para o e-mail labbdciferri@gmail.com, com o seguinte assunto: [Organização de Arquivos] Trabalho Prático 2016 – Turma X – Parte Y. Deve constar no corpo da mensagem o NUSP e nome de cada integrante do grupo. Não será atribuída nota ao aluno cujos dados não constarem no corpo da mensagem.

[2] Entregar pessoalmente a documentação externa impressa em horário e local definido na página da disciplina.

Critério de Correção

Critério de avaliação do trabalho. Na correção do trabalho, serão ponderados os seguintes aspectos.

- Qualidade da documentação (interna e externa) entregue. MAIOR PESO
- Corretude da execução do programa. MAIOR PESO
- Qualidade da interface.

Restrições adicionais sobre o critério de correção.

- A não execução de um programa devido a erros de compilação implica que a nota final da parte do trabalho será igual a zero (0).
- A ausência da documentação interna implica que haverá uma diminuição expressiva na nota do trabalho.
- A ausência da documentação externa implica que haverá uma diminuição expressiva na nota do trabalho.
- A inserção de palavras ofensivas nos arquivos e em qualquer outro material entregue implica que a nota final da parte do trabalho será igual a zero (0).
- Em caso de cola, as notas dos trabalhos envolvidos será igual a zero (0).
- Devem ser exibidos avisos ou mensagens de erro quando apropriado.

Critério de avaliação dos integrantes. Podem ser incluídas uma ou mais perguntas a respeito do trabalho na prova.

Data de Entrega do Trabalho

Nas datas especificadas na página da disciplina.

Bom Trabalho !