# 22667– Organização e Recuperação da Informação

Jander Moreira*

Última revisão: 1 de Agosto de 2018

## Conteúdo

**1 Introdução** **1**

**2 Módulos** **1**

2.0 Representação da informação . . . 1
2.1 Arquivos e armazenamento . . . . 2
2.2 Indexação e estruturas de indexação 2
2.3 Tabelas hash . . . . . 3
2.4 Métodos de ordenação . . . . . 3
2.5 Coleta de lixo . . . . . 3
2.6 Compressão de dados . . . . . 4

## 1 Introdução

Este é o planejamento para uma oferta da disciplina 22667 – Organização e Recuperação da Informação, com previsão de duração de 30 aulas de 1h40min cada uma.

## 2 Módulos

A disciplina é dividida em temas, caracterizados como módulos, indicados na Tabela I. Incorpora-se ao primeiro módulo a apresentação da disciplina.

### 2.0 Representação da informação

Apresentação da disciplina deve ter cerca de 30 minutos, abordando o contexto da disciplina, seu conteúdo geral, os critérios de avaliação e frequência, o formato da condução das aulas e demais atividades.

Este módulo explora o conceito geral de símbolos e sua formas de representação, com conceito de dados e informação. Leva, então, a uma revisão da representação binária de tipos de dados usuais, com inteiros com e sem sinal, valores com ponto flutuante, valores lógicos e cadeias de caracteres.

*Moreira, J. – Universidade Federal de São Carlos – Departamento de Computação – *Rodovia Washington Luis, km 235 – 13565-905 - São Carlos/SP – Brasil* – jander@dc.ufscar.br

Tabela I: Módulos da oferta, com número de minutos de dedicação em aula e quantidade de aulas de duas horas necessária.

| Módulo | Tema | Minutos | Aulas |
|---|---|---|---|
| 0 | Representação da informação | 100 | 1 |
| 1 | Arquivos e armazenamento | 400 | 4 |
| 2 | Indexação e estruturas de indexação | 600 | 6 |
| 3 | Tabelas hash | 400 | 4 |
| 4 | Métodos de ordenação | 500 | 5 |
| 5 | Coleta de lixo | 200 | 2 |
| 6 | Compressão de dados | 200 | 2 |
| | Avaliações | 300 | 3 |

As atividades deste módulo incluem a exposição do conteúdo e a realização de uma atividade em aula. São propostos, também, exercícios para estudo extra-classe.

O conteúdo do módulo e os tempos previstos estão apresentados na Tabela II.

Tabela II: Atividades e tempos estimados para o módulo de representação e armazenamento da informação.

| Tópico | Tempo (min) |
|---|---|
| Apresentação da disciplina | 30 |
| Aula expositiva | 40 |
| Atividade de aula: reforço do conceito de representação de dados | 30 |

A parte de representação da informação compreende os seguintes pontos:

- Apresentação dos conceitos de códigos: símbolos e regras de codificação;
- Revisão da representação binária para tipos de dados comuns;

Em particular, são abordados os conceitos:

- Símbolos e codificação básica;
- Representações usuais
  - Inteiros com e sem sinal;
  - Reais com ponto flutuante (visão geral do IEEE-754);
  - Cadeias de caracteres terminadas em sentinela e com prefixo de tamanho;
  - Lógicos individuais ou mapa de bits;
  - Representações alternativas, como BCD ou textual.

## 2.1 Arquivos e armazenamento

Uma visão geral sobre a tecnologia de dispositivos de armazenamento secundário é apresentada, caracterizando suas diferenças de espaço e desempenho de acesso. Também uma visão básica sobre o gerenciamento do espaço é apresentado com o conceito de sistemas de arquivos, sem detalhamento de sua estrutura. Em especial, é dada atenção ao modo de recuperação de dados em blocos (páginas, *clusters*) e sua relevância para a recuperação dos dados.

A manipulação de arquivos é abordada, com revisão geral dos conceitos de acesso por programação (abertura, fechamento, leitura, escrita, posicionamento no fluxo de dados).

A expectativa de atividades do módulo é apresentada na Tabela III.

Tabela III: Atividades e tempos estimados para o módulo de manipulação de arquivos e armazenamento.

| **Tópico** | **Tempo (min)** |
|---|---|
| Aula expositiva | 100 |
| Atividade de aula: programação com arquivos | 100 |
| Atividade de aula: campos, registros e operações | 100 |
| Atividade de aula: reuso de espaço | 100 |

Este módulo cobre:

- Noções de armazenamento externo: mídias e organização de dados
- Manipulação de arquivos
- Representação de campos, registros e arquivos, com manipulação básica e reuso de espaço

De forma mais detalhada:

- Armazenamento externo
  - Visão geral de dispositivos: fitas, discos rígidos e estado sólido
  - Visão geral de sistemas de arquivos e recuperação de dados em blocos
- Conceito e manipulação de arquivos
  - Tipos texto e binários
  - Acesso sequencial e aleatório
  - Operações de acesso a arquivos: abertura, fechamento, leitura e escrita
- Conceito de campos e registros
  - Representações de campos de tamanho fixo e variável
  - Representações de registros de tamanho fixo e variável, bem como campos fixos e variáveis
  - Remoção lógica de registros e reaproveitamento de espaço

## 2.2 Indexação e estruturas de indexação

O tema de índices para arquivos de dados é coberto por este módulo.

Inicialmente o assunto da indexação é abordado, cobrindo-se as diferentes formas básicas de estruturá-la. São introduzidos os conceitos de índices lineares e em múltiplos níveis, bem como de índices primários e secundários.

As estruturas básicas de indexação são cobertas juntamente com suas operações, cobrindo os mecanismos de inserção e remoção.

A principal estrutura de árvore para indexação é apresentada: árvore B. Os algoritmos de busca, inserção e remoção são abordados com profundidade.

Uma visão complementar sobre as estruturas de árvore é apresentada na forma de árvores B+ (ou B*), comparativamente às árvores B convencionais.

Uma visão sobre dados não convencionais e estruturas para armazená-los ou indexá-los é apresentada. Àrvores como quad-trees, k-d e métricas são introduzidas.

A Tabela IV apresenta as atividades previstas para este módulo, juntamente com sua duração.

Este módulo cobre:

- Indexação;
- Estruturas de dados para indexação.

Com uma especificação mais detalhada:

- Conceitos de índices
  - Lineares e multiníveis;
  - Primários e secundários;
- Estruturas de índices e operações
  - Operações de inserção, busca e remoção;
  - Listas invertidas;
- Árvores multi-caminhos

Tabela IV: Atividades e tempos estimados para o módulo de indexação e estruturas de indexação.

| Tópico | Tempo (min) |
|---|---|
| Aula expositiva | 100 |
| Atividade de aula: indexação, estruturas básicas | 100 |
| Atividade de aula: indexação, estruturas básicas | 100 |
| Aula expositiva | 100 |
| Atividade de aula: árvores B e operações de busca e inserção | 100 |
| Atividade de aula: árvores B e operação de remoção | 100 |

  - B: estrutura e operações;
  - B+: visão geral;
  - Visão geral sobre dados não convencionais.

### 2.3 Tabelas hash

A abordagem sobre tabelas hash apresenta o conceito de tabelas de endereçamento direto e transpõe o conceito para tabelas hash. Com o uso do hash, há a possibilidade de colisão e se apresentam as possibilidades de tratamento: encadeamento e endereçamento aberto. Ambas são avaliadas quanto a operação e desempenho.

A questão se necessitar uma função hash é abordada sob a perspectiva de desempenho da função e do qualidade do espalhamento proporcionado. São coberto alguns do métodos de estruturação de funções hash, como divisão, multiplicação, dobramento e meio-quadrado, por exemplo. Também é dada uma visão geral sobre o conceito de hash universal, suas características e projeto.

Noções sobre hash perfeito são apresentadas.

Na TabelaV estão as atividades previstas para este módulo.

A abordagem sobre hash envolve:

- Tabelas de endereçamento direto
- Tabelas hash: funções, tratamento de colisão
- Hash perfeito

Mais especificamente:

- Tabelas de endereçamento direto
- Tabelas hash
- Funções hash
  - Métodos de criação de funções
  - Hash universal
- Tratamento de colisões
  - Encadeamento

Tabela V: Atividades e tempos estimados para o módulo de tabelas hash.

| Tópico | Tempo (min) |
|---|---|
| Aula expositiva | 100 |
| Atividade de aula: hash e tratamento por encadeamento | 100 |
| Atividade de aula: funções hash | 100 |
| Atividade de aula: hash e tratamento por endereçamento aberto | 100 |

  - Endereçamento aberto
- Hash perfeito

### 2.4 Métodos de ordenação

A ordenação de dados é o tema deste módulo. São cobertas as principais estratégias de ordenação tanto para memória principal quanto secundária.

Para a memória principal são vistos os algoritmos elementares: inserção, seleção e bubblesort. Para cada um deles é apresentada uma noção sobre seu desempenho. Além destes, são apresentados os métodos não elementares cujo desempenho é notadamente superior: shellsort, heapsort e quicksort. Cada um deles é observado sobre seu funcionamento e desempenho.

Para a memória secundária é salientada a necessidade de diferenciação em relação à ordenação em memória primária. Os processos de intercalação multi-vias é introduzido, sendo vistas as técnicas de intercalação simples, balanceada e polifásica.

A Tabela VI apresenta as atividades previstas para este módulo, juntamente com sua duração.

Este módulo cobre:

- Ordenação em memória primária;
- Ordenação em memória secundária.

De forma mais detalhada:

- Ordenação interna
  - Métodos elementares e seus desempenhos: inserção, seleção e bubblesort;
  - Métodos não elementares e seus desempenhos: shellsort, heapsort e quicksort;
- Ordenação externa
  - Noções de processamento cossequencial;
  - Ordenação por intercalação: básica, balanceada e polifásica;

### 2.5 Coleta de lixo

A coleta de lixo é abordada dentro do contexto de gerenciamento de memória.

Tabela VI: Atividades e tempos estimados para o módulo de representação e armazenamento da informação.

| Tópico | Tempo (min) |
|---|---|
| Aula expositiva | 100 |
| Atividade de aula: métodos elementares | 60 |
| Atividade de aula: shellsort | 40 |
| Atividade de aula: heapsort e quicksort | 100 |
| Atividade de aula: intercalação | 100 |
| Atividade de aula: intercalação balanceada | 50 |
| Atividade de aula: intercalação polifásica | 50 |

O gerenciamento de memória é apresentado de forma geral, abordando os métodos de ajuste sequencial e não sequencial.

O gerenciamento automático de memória é introduzido e as diferentes técnicas de coleta de lixo são analisadas: marcar e colar e cópia. Uma visão sobre coleta de lixo incremental é apresentada.

Na TabelaVIII estão as atividades previstas para este módulo.

Tabela VII: Atividades e tempos estimados para o módulo de coleta de lixo.

| Tópico | Tempo (min) |
|---|---|
| Aula expositiva | 100 |
| Atividade de aula: gerenciamento sequencial e não sequencial | 30 |
| Atividade de aula: coleta de lixo | 70 |

Este módulo cobre:

- Visão geral sobre gerenciamento da memória principal;
- Coleta de lixo.

Com uma especificação mais detalhada:

- Uso da memória
- Métodos de ajuste sequencial e não sequencial
- Abordagens para coleta de lixo
  - Marcar e trocar
  - Cópia
  - Incremental

## 2.6 Compressão de dados

A abordagem sobre compressão de dados retoma o conceito de representação da informação e o uso de símbolos, tornando a compressão um processo reversível de codificação com códigos de comprimento menor de bits.

Dois métodos de compressão sem perdas são abordados: Huffman e LZW. As semelhanças e diferenças são observadas.

Na TabelaVIII estão as atividades previstas para este módulo.

Tabela VIII: Atividades e tempos estimados para o módulo de coleta de lixo.

| Tópico | Tempo (min) |
|---|---|
| Aula expositiva | 100 |
| Atividade de aula: gerenciamento sequencial e não sequencial | 40 |
| Atividade de aula: coleta de lixo | 60 |

O módulo de compressão de dados aborda:

- Revisão geral do conceito de símbolos e representação;
- Técnicas de compressão de dados sem perda de informação.

Com uma especificação mais detalhada:

- Visão geral de símbolos e representação;
- Métodos de compressão
  - Codificação de Huffman
  - LZW