Classificação e Pesquisa de Dados

Aula 01
Apresentação da Disciplina; Introdução à Classificação; Definições

INF01124

INF01124 - CPD

Renata Galante
galante@inf.ufrgs.br
sala 221 – prédio 43.424
ramal 7746

moodle.inf.ufrgs.br
senha: **inf01124b**

Classificação de Dados

UFRGS INF01124

Esta disciplina

- ALGORÍTMOS E PROGRAMAÇÃO – CIC
- FUNDAMENTOS DE ALGORITMOS
- ESTRUTURAS DE DADOS
  - ALGORÍTMOS E PROGRAMAÇÃO
  - FUNDAMENTOS DE ALGORITMOS
- **CLASSIFICAÇÃO E PESQUISA DE DADOS**
  - ESTRUTURAS DE DADOS
- TÉCNICAS DE CONSTRUÇÃO DE PROGRAMAS
  - CLASSIFICAÇÃO E PESQUISA DE DADOS
- FUNDAMENTOS DE BANCO DE DADOS
  - CLASSIFICAÇÃO E PESQUISA DE DADOS

Objetivos da disciplina

- Capacitar o aluno para seleção e análise de:
  - Algoritmos para classificação de dados
  - Algoritmos para pesquisa de dados
  - Técnicas de organização e de compactação de arquivos
- Capacitar os alunos para a disciplina de Banco de Dados
- Considera estruturas de dados em memória e em disco

Súmula

1. Métodos de Classificação de Dados
2. Introdução à Análise de Complexidade de Algoritmos
3. Métodos de Armazenamento e Pesquisa de Dados em Tabelas
4. Técnicas de Organização de Arquivos
5. Técnicas de Compactação de Arquivos

Classificação (*Sorting*)

- Processo de organizar itens em ordem (de)crescente, segundo algum critério
- Também chamado de ordenação
- Aplicações da Classificação:
  - Preparação de dados para facilitar pesquisas futuras
    - Exemplo: dicionários e listas telefônicas
  - Agrupar itens que apresentam mesmos valores
    - Para eliminação de elementos repetidos
  - Identificação de itens presentes em mais de um arquivo
    - Para combinação de dados presentes nos diferentes arquivos;
    - Para consolidação dos vários arquivos em um único

Definições

- Sejam $R_1, R_2, \ldots, R_n$, n itens (chamados registros)
- Cada registro R_i , é formado por uma chave C_i e por outros dados ditos satélites
- A ordenação dos registros é feita definindo-se uma relação de ordem "<" sobre os valores das chaves
- O objetivo da ordenação é determinar uma permutação dos índices $1 \leq i_1, i_2, \ldots, i_n \leq n$ das chaves, tal que
- Um conjunto de registros é chamado de arquivo

$$C_{i_1} \leq C_{i_2} \leq \ldots \leq C_{i_n}$$

Relação de Ordem

- Uma relação de ordem "**<**" (leia-se: precede) deve satisfazer as seguintes condições para quaisquer valores ***a***, ***b*** e ***c***:

  (i) Uma e somente uma das seguintes possibilidades é verdadeira (lei da tricotomia):

  a* < *b, ***a* = *b*** ou ***b* < *a***

  (ii) Se ***a* < *b*** e ***b* < *c***, então ***a* < *c*** (transitividade)

- As propriedades (i) e (ii) definem o conceito de ordem linear ou ordem total

Mais Definições

- Um algoritmo de classificação é dito **estável**, se ele preserva a ordem relativa original dos registros com mesmo valor de chave
- Já a classificação é **local** quando for feita sobre a mesma área física onde se encontram as chaves (não há necessidade de memória extra).
- Algoritmos de ordenação podem ser classificados como **internos** (todos os registros mantidos em RAM) ou **externos** (utilizando dispositivos eletro-mecânicos de armazenagem da dados).

Memória e disco

- O problema surge quando o número de registros a serem classificados é maior do que a quantidade que pode ser mantida em memória principal.
- Matrizes (*arrays*) em memória e arquivos em disco
- As estruturas de dados devem ser armazenadas em dispositivos de armazenamento lentos.
- Há diferença de duas ou três ordens de grandeza entre o acesso a memória e o acesso ao disco
- A *memória* tem tempo de acesso constante e igual para cada posição.
- O *disco* tem tempo de acesso variável e dependente da região onde estão localizados os dados.

Qual é a diferença?

- Em *memória*
  - acesso é direto às estruturas de dados (qualquer palavra da memória é diretamente acessível).
- Em *disco*
  - é preciso buscar os elementos de dados de um dispositivo através de ações mecânicas: (tempo de acesso = *seek* + demora rotacional + transferência)
- Devemos considerar as diferenças quando da especificação dos algoritmos.

Formas de Representação do Resultado

- Reorganização Física
- Encadeamento
- Vetor Indireto de Ordenação (VIO) ou Índice

Classificação de Dados com Reorganização Física

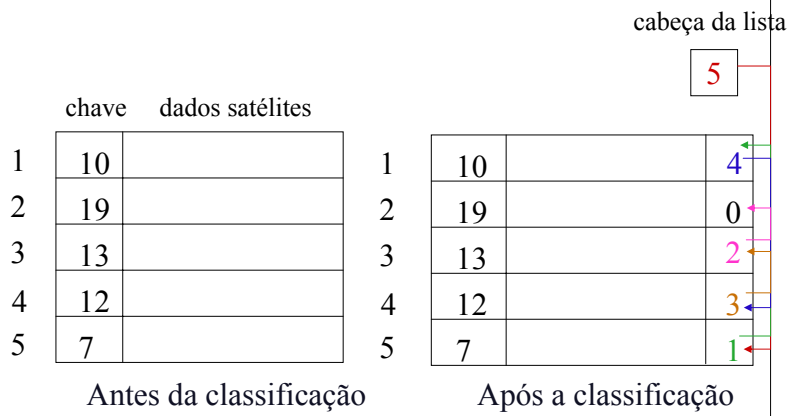
	chave	dados satélites
1	10	
2	19	
3	13	
4	12	
5	7	

Antes da classificação

	chave	dados satélites
1	7	
2	10	
3	12	
4	13	
5	19	

Após a classificação

Classificação de Dados através de Encadeamento



- Permite somente acesso seqüencial aos registros ordenados!

Classificação de Dados por Vetor Indireto de Ordenação (VIO) ou Índice

	chave	dados satélites
1	10	
2	19	
3	13	
4	12	
5	7	

Índice

	chave	
1	7	5
2	10	1
3	12	4
4	13	3
5	19	2

- Acesso seqüencial ou acesso por pesquisa binária, mas sempre por via indireta

Arquivo ordenado simultaneamente por várias chaves

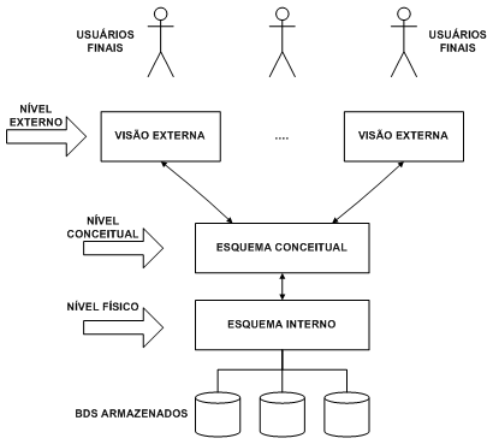
Arquivo Original

	Chave 1	Chave 2	Chave 3	
1	12	4	10	
2	25	18	21	
3	3	14	2	
4	9	20	13	
5	17	1	6	

Arquivo ordenado simultaneamente por várias chaves

	Chave 1	Chave 2	Chave 3		VIO Chave 2	VIO Chave 3
1	3	14	2		4	1
2	9	20	13		3	4
3	12	4	10		1	3
4	17	1	6		5	2
5	25	18	21		2	5

Banco de Dados



Objetivos da disciplina

- Capacitar o aluno para seleção e análise de:
  - Algoritmos para classificação de dados
  - Algoritmos para pesquisa de dados
  - Técnicas de organização e de compactação de arquivos
- Capacitar os alunos para a disciplina de Banco de Dados
- Considera estruturas de dados em memória e em disco

Súmula

1. Métodos de Classificação de Dados
2. Introdução à Análise de Complexidade de Algoritmos
3. Métodos de Armazenamento e Pesquisa de Dados em Tabelas
4. Técnicas de Organização de Arquivos
5. Técnicas de Compactação de Arquivos

Bibliografia

AZEREDO, P. A. **Métodos de Classificação de Dados e Análise de suas Complexidades**. Editora Campus, RJ, 1995.
FURTADO, A. L.; SANTOS, C. S. dos. **Organização de Banco de Dados**. Editora Campus, Rio de Janeiro, 1988.
SANTOS, C. S.; AZEREDO, P. A. **Tabelas. Organização e Pesquisa**. Série Livros Didáticos, Editora Sagra Luzzato, Porto Alegre, 2001.

Leitura complementar (fortemente recomendada):

CORMEN, T.; LEISERSON, C.; RIVEST, R. **Introduction to Algorithms**. The MIT Press. Cambridge, Massachusetts, 1990.
KNUTH, D. **The Art of Computer Programming**: Sorting and Searching. Vol. 2. Addison-Wesley, Reading, Mass, 1973.
SZWARCFITER, Jayme L.; MARKENZON, Lilian. **Estrutura de Dados e seus algoritmos**. Rio de Janeiro: LTC, 1994.
Material disponível na Web principalmente o material disponível na Wikipedia e suas referências (o aluno deve desenvolver espírito crítico no estudo deste material devido a edição cooperativa do mesmo).

Avaliação

- Serão realizadas duas provas (P1 e P2), um trabalho prático final (TF) e exercícios/tarefas (LET). Também será considerada a participação em aula.
- A média geral (MG) é a média ponderada dos graus obtidos na provas e trabalho acima referidos, e será calculada pela fórmula:

  MG = 0,35*P1 + 0,35*P2 + 0,20*TF + 0,10*LET
- A conversão da MG para conceitos é feita por meio da seguinte tabela:

  9,0 ≤ **MG** = 10,0 : conceito **A** (aprovado)
  7,5 ≤ **MG** < 9,0 : conceito **B** (aprovado)
  6,0 ≤ **MG** < 7,5 : conceito **C** (aprovado)
  4,0 ≤ **MG** < 6,0 : sem conceito (**recuperação**)
  0,0 ≤ **MG** < 4,0 : conceito **D** (reprovado)
  Faltas > 25% : conceito **FF** (reprovado)

Recuperação - I

1. Somente serão calculadas as médias gerais daqueles alunos que tiverem, ao longo do semestre, obtido um índice de freqüência às aulas igual ou superior a 75% das aulas previstas. Aos que não satisfizerem este requisito, será atribuído o conceito FF (Falta de Freqüência)
2. Para poder realizar a prova de recuperação, o aluno deve ter realizado as duas provas (P1 e P2), ter entregado o trabalho final (TF) e ter realizado mais de 2/3 das listas de exercícios e tarefas (LET). Além disso deverá ter nota igual ou superior a 6,0 em pelo menos uma das duas provas
3. Os que não se enquadrarem nessa situação receberão conceito D

Recuperação - II

1. Serão considerados aprovados na recuperação os alunos que obtiverem um aproveitamento de no mínimo 60% da prova. A estes será atribuído o conceito C. Aos demais, o conceito D.
2. Não há recuperação das provas P1 e P2 por não comparecimento, exceto nos casos previstos na legislação (saúde, parto, serviço militar, convocação judicial, luto etc. devidamente comprovados). Nesse caso, o aluno deverá fazer a prova de recuperação.

Exercício

- Escreva um algoritmo para classificar um conjunto de **5 números** em ordem **crescente**