

Algoritmos e Estruturas de Dados II

Aula 29 – Ordenação Externa - Intercalação usando Seleção por Substituição e Polifásica

Prof. Luciano A. Digiampietri
digiampietri@usp.br
@digiampietri

2024

Intercalação via Seleção por Substituição

Intercalação Balanceada de f caminhos: dá para melhorar?

- Obs: O método de intercalação que vimos até agora é chamado de intercalação balanceada, pois distribui de forma balanceada as corridas pelas vias.
- Lembrando que há duas fases: ordenação (para geração de corridas iniciais (ordenadas) e intercalação
- Número de passos de intercalação $P(N) = \lceil (\log_f \lceil (N/M) \rceil) \rceil$, sendo $\lceil (N/M) \rceil$ o número de corridas iniciais

Racional para um novo método: se diminuirmos o número de corridas (gerando corridas mais longas), podemos diminuir o número de passos.

Intercalação via Seleção por Substituição

Intercalação Balanceada de f caminhos: dá para melhorar?

- Obs: O método de intercalação que vimos até agora é chamado de intercalação balanceada, pois distribui de forma balanceada as corridas pelas vias.
- Lembrando que há duas fases: ordenação (para geração de corridas iniciais (ordenadas) e intercalação
- Número de passos de intercalação $P(N) = \lceil (\log_f \lceil (N/M) \rceil) \rceil$, sendo $\lceil (N/M) \rceil$ o número de corridas iniciais

Enquanto existirem chaves que podem entrar na intercalação (maiores ou iguais às anteriores) permanecemos na mesma corrida.

Ordenação - Seleção por Substituição

Implementação por meio de Seleção por Substituição

- A implementação do método de intercalação balanceada pode ser realizada utilizando filas de prioridades.
- As duas fases do método podem ser implementadas de forma eficiente e elegante.
- Operações básicas para formar blocos ordenados:
 - Obter o menor dentre os registros presentes na memória interna.
 - Substituí-lo pelo próximo registro da fita de entrada.
- Estrutura ideal para implementar as operações: *heap*.
- Operação de substituição:
 - Retirar o menor item da fila de prioridades.
 - Colocar um novo item no seu lugar.
 - Reconstituir a propriedade do *heap*.

(ZIVIANI, 2010)

Ordenação - Seleção por Substituição

Geração das corridas iniciais: conseguiremos gerar corridas iniciais maiores que M , usando um heap de tamanho M) - sendo M o tamanho da memória disponível

Ordenação - Seleção por Substituição

Algoritmo

1. Inserir m elementos do arquivo original de entrada na fila de prioridades
2. Substituir o menor item da fila de prioridades pelo próximo item do arquivo
3. Se esse próximo item é menor do que o que saiu:
 - Marque-o como membro da próxima corrida
 - Trate-o como maior que todos os itens da corrida atual
4. Se um item marcado for para o topo da fila de prioridades então:
 - A corrida atual é encerrada
 - Uma nova corrida ordenada é iniciada

Ordenação Externa - Seleção por Substituição

Entrada: INTERCALACAOBALANCEADA

H 1:

E 2:

A 3:

P

Sai:

Entra:

Ordenação Externa - Seleção por Substituição

Entrada: INTERCALACAOBALANCEADA

H	1:	I
E	2:	N
A	3:	T
P		

Sai:

Entra:

Ordenação Externa - Seleção por Substituição

Entrada: INTERCALACAOBALANCEADA

H 1: N
E 2: E*
A 3: T
P

Sai: I

Entra: E

Ordenação Externa - Seleção por Substituição

Entrada: INTERCALACAOBALANCEADA

H 1: R
E 2: E*
A 3: C*
P

Sai: I N

Entra: E R

Ordenação Externa - Seleção por Substituição

Entrada: INTERCALACAOBALANCEADA

H 1: T
E 2: E*
A 3: C*
P

Sai: I N R

Entra: E R C

Ordenação Externa - Seleção por Substituição

Entrada: INTERCALACAOBALANCEADA

H 1: A*

E 2: E*

A 3: C*

P

Sai: I N R T

Entra: E R C A

Ordenação Externa - Seleção por Substituição

Entrada: INTERCALACAOBALANCEADA

H	1:	A
E	2:	E
A	3:	C
P		

Sai: I N R T

Entra: E R C A

Primeira corrida: INRT

Ordenação Externa - Seleção por Substituição

Entrada: INTERCALACAOBALANCEADA

H 1: A
E 2: E
A 3: C
P

Sai: I N R T A C E L A A B C L O A A C E N A A D

Entra: E R C A L A C A O B A L A N C E A D A

Corridas: INRTACELAABCL O AACEN AAD

Ordenação Externa - Seleção por Substituição

Entrada: INTERCALACAOBALANCEADA

H 1: A
E 2: E
A 3: C
P

Sai: I N R T A C E L A A B C L O A A C E N A A D
Entra: E R C A L A C A O B A L A N C E A D A

Corridas: INRTACEL AABCL O AACEN AAD

Cinco corridas ao invés de oito.

Ordenação - Seleção por Substituição

Fases da Intercalação:

Algoritmo:

Ideia: Há f fitas para intercalar, cada uma com no máximo C corridas. Vou manter um heap de tamanho máximo f , contendo o menor elemento de cada fita (mais precisamente, da atual corrida de cada fita) Para cada conjunto de i -ésimas corridas das fitas ($i = 1, \dots, C$)

1. Monte uma fila de prioridades de tamanho f
2. Enquanto não estiver vazia:
 - Substitua o menor item k (do topo da fila de prioridades) pelo próximo item da mesma corrida da mesma fita do item k (se houver)
 - Imprima na fita de saída o item k substituído

Fases da Intercalação

fita 1:	I	N	R	T		A	A	C	E	N
fita 2:	A	C	E	L		A	A	D		
fita 3:	A	A	B	C	L	O				
fita 4:										

Fases da Intercalação

fita 1:	I	N	R	T		A	A	C	E	N
fita 2:	A	C	E	L		A	A	D		
fita 3:	A	A	B	C	L	O				
fita 4:	A									

Fases da Intercalação

fita 1:	INRT	AACEN
fita 2:	ACEL	AAD
fita 3:	AABCL0	
fita 4:	AA	

Fases da Intercalação

fita 1:	INRT	AACEN
fita 2:	ACEL	AAD
fita 3:	AABCLO	
fita 4:	AAA	

Fases da Intercalação

fita 1:	INRT	AACEN
fita 2:	ACEL	AAD
fita 3:	AABCL	O
fita 4:	AAAB	

Fases da Intercalação

fita 1:	INRT	AACEN
fita 2:	ACEL	AAD
fita 3:	AABCL	O
fita 4:	AAABC	

Fases da Intercalação

fita 1:	INRT	AACEN
fita 2:	ACEL	AAD
fita 3:	AABCL	O
fita 4:	AAABCC	

Fases da Intercalação

fita 1:	INRT	AACEN
fita 2:	ACEL	AAD
fita 3:	AABCL	O
fita 4:	AAABCCE	

Fases da Intercalação

fita 1:	INRT	AACEN
fita 2:	ACEL	AAD
fita 3:	AABCL	O
fita 4:	AAABCCEI	

Fases da Intercalação

fita 1:	INRT	AACEN
fita 2:	ACEL	AAD
fita 3:	AABCL0	
fita 4:	AAABCCEIL	

Fases da Intercalação

fita 1:	INRT	AACEN
fita 2:	ACEL	AAD
fita 3:	AABCL0	
fita 4:	AAABCCEILL	

Fases da Intercalação

fita 1: INRT AACEN
fita 2: ACEL AAD
fita 3: AABCL0
fita 4: AAABCCEILLN

Fases da Intercalação

fita 1: INRT AACEN
fita 2: ACEL AAD
fita 3: AABCL0
fita 4: AAABCCEILLNO

Fases da Intercalação

fita 1: INR**T** AACEN
fita 2: ACEL AAD
fita 3: AABCL0
fita 4: **AAABCCEILLNOR**

Fases da Intercalação

fita 1: INRT AACEN
fita 2: ACEL AAD
fita 3: AABCL0
fita 4: AAABCCEILLNORT

Fases da Intercalação

fita 1: AACEN

fita 2: AAD

fita 3:

fita 4: AAABCCEILLNORT

Fases da Intercalação

fita 1: AACEN

fita 2: AAD

fita 3:

fita 4: AAABCCEILLNORT AAAACDEN

Fases da Intercalação

fita 1: AAAACDEN

fita 2:

fita 3:

fita 4: AAABCCEILLNORT

Fases da Intercalação

fita 1: AAAACDEN

fita 2: AAAAAABCCCDEEILLNORT

fita 3:

fita 4: AAABCCEILLNORT

Ordenação - Seleção por Substituição

Vantagens

- Número de comparações para achar o menor item (fase de intercalação)
 - Intercalação balanceada de f caminhos: $f-1$
 - Seleção por substituição:
 - $\log_2(f)$ (usando heap binário)
- Corridas iniciais de tamanho médio 2^*M : número menor de passadas sobre o arquivo para ordená-lo (intercalações)

Vantajoso quando f é grande (maior ou igual a 8)

Ordenação - Seleção por Substituição

Considerações Práticas

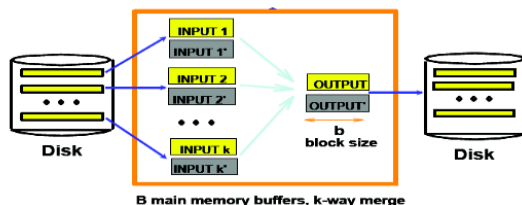
- As operações de entrada e saída de dados devem ser implementadas eficientemente.
- Deve-se procurar realizar a leitura, a escrita e o processamento interno dos dados de forma simultânea.
- Os computadores de maior porte possuem uma ou mais unidades independentes para processamento de entrada e saída.
- Assim, pode-se realizar processamento e operações de E/S simultaneamente.

(ZIVIANI, 2010)

Ordenação - Seleção por Substituição

Considerações Práticas

- Técnica para obter superposição de E/S e processamento interno:
 - Utilize $2f$ áreas de entrada e $2f$ de saída. (ou $2f$, como o desenho abaixo)
 - Para cada unidade de entrada ou saída, utiliza-se duas áreas de armazenamento:
 1. Uma para uso do processador central
 2. Outra para uso do processador de entrada ou saída.
 - Para entrada, o processador central usa uma das duas áreas enquanto a unidade de entrada está preenchendo a outra área.
 - Depois a utilização das áreas é invertida entre o processador de entrada e o processador central.
 - Para saída, a mesma técnica é utilizada.



Vantagem: processa mais rápido

Desvantagem: diminui f
aumenta o número de
passos de intercalação

(ZIVIANI, 2010)

Ordenação - Seleção por Substituição

Considerações Práticas

- Escolha da ordem de intercalação f :
 - Para fitas magnéticas:
 - * f deve ser igual ao número de unidades de fita disponíveis menos um.
 - * A fase de intercalação usa f fitas de entrada e uma fita de saída.
 - * O número de fitas de entrada deve ser no mínimo dois.
 - Para discos magnéticos:
 - * O mesmo raciocínio acima é válido.
 - * O acesso seqüencial é mais eficiente.
 - Sedgwick (1988) sugere considerar f grande o suficiente para completar a ordenação em poucos passos.
 - Porém, a melhor escolha para f depende de vários parâmetros relacionados com o sistema de computação disponível.

(ZIVIANI, 2010)

Ordenação - Intercalação Polifásica

Intercalação Polifásica

- Problema com a intercalação balanceada de vários caminhos:
 - Necessita de um grande número de fitas.
 - Faz várias leituras e escritas entre as fitas envolvidas.
 - Para uma intercalação balanceada de f caminhos são necessárias $2f$ fitas.
 - Alternativamente, pode-se copiar o arquivo quase todo de uma única fita de saída para f fitas de entrada.
 - Isso reduz o número de fitas para $f + 1$.
 - Porém, há um custo de uma cópia adicional do arquivo.
- Solução:
 - **Intercalação polifásica.** (para otimizar o uso das fitas)

(ZIVIANI, 2010)

Ordenação - Intercalação Polifásica

Intercalação Polifásica

- As corridas ordenadas são distribuídas de forma desigual entre as fitas disponíveis.
- Uma fita é deixada livre. (fará o papel de fita de saída)
- Em seguida, a intercalação de corridas ordenadas é executada até que uma das fitas esvazie. (que se tornará a nova fita de saída)
- Neste ponto, a antiga fita de saída troca de papel com a fita de entrada.

Isso elimina a necessidade de redistribuição do conteúdo da fita de saída

(ZIVIANI, 2010)

Ordenação - Intercalação Polifásica

Intercalação Polifásica

- Exemplo: (posso combinar seleção por substituição com intercalação polifásica)
 - Corridas ordenadas obtidas por meio de seleção por substituição:

fita 1:	<i>I N R T</i>	<i>A C E L</i>	<i>A A B C L O</i>
fita 2:	<i>A A C E N</i>	<i>A A D</i>	
fita 3:			

Fita 3 é de saída

- Configuração após uma intercalação-de-2-caminhos das fitas 1 e 2 para a fita 3:

fita 1:	<i>A A B C L O</i>	
fita 2:		
fita 3:	<i>A A C E I N N R T</i>	<i>A A A C D E L</i>

Fita 2 ficou vazia
(então se torna a
nova fita de saída)

(ZIVIANI, 2010)

Ordenação - Intercalação Polifásica

Intercalação Polifásica

- Exemplo:
 - Depois da intercalação-de-2-caminhos das fitas 1 e 3 para a fita 2:

fita 1:
fita 2: *AAAABCCCEILNNORT*
fita 3: *AAACDEL*

Fita 1 ficou vazia
(então se torna a
nova fita de saída)

- Finalmente:

fita 1: *AAAAAABCCCEEILLNNORT*
fita 2: *AAAAAABCCCEEILLNNORT*
fita 3: *AAAAAABCCCEEILLNNORT*

- A intercalação é realizada em muitas fases.
- As fases não envolvem todas as corridas.
- Nenhuma cópia direta entre fitas é realizada.

(ZIVIANI, 2010)

Ordenação - Intercalação Polifásica

Intercalação Polifásica

- A implementação da intercalação polifásica é simples.
- A parte mais delicada está na distribuição inicial das corridas ordenadas entre as fitas.
- Distribuição das corridas nas diversas etapas do exemplo:

fita 1	fita 2	fita 3	Total
3	2	<u>0</u>	5
1	<u>0</u>	2	3
<u>0</u>	1	1	2
1	0	0	1

(fita de saída sublinhada de vermelho)

Estratégia: planejar de baixo para cima:

- Última linha: escolha uma fita para ter a corrida final – ela recebe 1 e o restante 0
- A cada linha i , de baixo para cima:
 - Considere o maior valor v da linha i ;
 - Na linha $i-1$, zere a fita que contém o valor v e adicione v ao valores das demais fitas
- Se no final o número total de corridas calculado for maior que o real, considere a diferença como corridas “dummy” (que não interferem na intercalação)

(ZIVIANI, 2010)

Considerações Finais:

- Fatores importantes para uma rápida ordenação externa:
 - Crie as corridas iniciais o mais longas que for possível
 - Sobreponha processamento do algoritmo, leitura e escrita o quanto puder
 - Use o máximo de memória possível
 - Se possível, use o maior número de discos que puder para ter acesso sequencial

Referência

- Slides baseados no material da profa. Arianne Machado Lima - ACH2024
- ELMARIS, R.; NAVATHE, S. B. Fundamentals of Database Systems. 4 ed. Ed. Pearson-Addison Wesley. Cap 13.8. 4 ed. Pearson. 2004
- RAMAKRISHNAN, R.; GEHRKE, J. Database Management Systems. 3 ed. Ed. McGraw-Hill. 2002 Cap 13
- ZIVIANI, N. Projeto de Algoritmos. Com Implementações em Pascal e C. Cengage. 3ª Edição, 2010.

Algoritmos e Estruturas de Dados II

Aula 29 – Ordenação Externa - Intercalação usando Seleção por Substituição e Polifásica

Prof. Luciano A. Digiampietri
digiampietri@usp.br
@digiampietri

2024