Universidade Federal da Bahia (UFBA)
Instituto de Computação (IME)
Departamento de Ciência da Computação (DCC)
MATA 54 - Estrutura de Dados e Algoritmos II – 2023.2 – Prof. George Lima

Estudo experimental do método *hashing* linear (HL)
18/09/2023

Este trabalho tem como objetivo consolidar conceitos sobre o método *hashing linear* (HL) [1, 2] através de sua avaliação experimental.

O relatório deve abordar o seguinte conteúdo:

- Descrição resumida do método.
- Descrição dos experimentos, com justificativas.
- Apresentação dos resultados e análise experimental.

1 Análise experimental

Serão objetos de avaliação experimental o comportamento do método HL em função dos seus parâmetros de configuração. O método HL a ser avaliado é o mesmo apresentado em sala. Se alguma variação deste for considerada, solicita-se que esta seja explicada no relatório.

Os parâmetros de configuração a serem considerados são os seguintes:

- Tamanho da página de dados $p \in \{1, 5, 10, 20, 50\}$.
- Fator de carga máximo $\alpha^{\max} \in \{0.2, 0.3, \ldots, 0.9\}$.

Deseja-se verificar como o método HL se comporta em função do tamanho da página (p) e do fator de carga máximo ($\alpha_{\max}$). Para tanto, solicita-se a inclusão simulada de chaves primárias, geradas aleatoriamente. O número de chaves aleatórias (distintas) a serem consideradas será $n = 1000 \times p$. Recomenda-se repetir, para cada valor de n, ao menos 10 vezes o experimento. Os resultados exibidos para cada n deve ser, portanto, uma média destas repetições.

O desempenho do método HL deverá ser avaliado de acordo com os seguintes fatores:

- Desempenho quanto ao espaço. Deseja-se verificar o número de páginas alocadas em relação ao necessário. Para tanto, deve-se medir o fator de carga médio e o número de páginas adicionais nas diversas listas, representados por $\alpha^{\text{médio}}$ e p^*, respectivamente, definidos como

$$\alpha^{\text{médio}} = \frac{\sum \text{espaço ocupado em cada página}}{\text{espaço total alocado}}$$

  e

$$p^* = \frac{\sum \text{número de páginas em cada lista}}{\text{número total de listas}}$$

- Desempenho quanto ao número médio de acessos. Nesta avaliação, deve-se medir o número médio de acessos para recuperar $k = \lceil 0.2n \rceil$ chaves, considerando busca com (C) e sem (S) sucesso. Para se calcular C deve-se escolher aleatoriamente k chaves, das n já incluídas. O cálculo de S deverá considerar k chaves geradas aleatoriamente e distintas daquelas já incluídas.

Mais especificamente, considerando que K^C e K^S são os conjuntos de chaves para o cálculo de C e S, define-se tais valores como

$$C = \frac{\sum_{k_i \in K^C} \#\text{ acessos para recuperar } k_i}{k}$$

e

$$S = \frac{\sum_{k_i \in K^S} \#\text{ acessos para recuperar } k_i}{k}$$

- Desempenho durante a inclusão dos n registros. O objetivo aqui é saber como o método HL se comporta à medida que o número de registros armazenados cresce. Para tanto, sugere-se fixar o valor de $p = 10$, e $\alpha_{\max} = 0.85$. Ao se incluir cada um dos $n = 10.000$ registros, deve-se computar, para $i = 1, 2, \ldots, n$,

$$\alpha^{\text{médio}}(i) = \frac{\sum \text{espaço ocupado em cada página}}{\text{espaço total alocado}},$$

$$p^*(i) = \frac{\sum \text{número de páginas em cada lista}}{\text{número total de listas}}$$

e

$$L^{\max}(i) = \text{número de páginas na maior lista}$$

2 Apresentação dos resultados

Os resultados devem ser apresentados, preferencialmente, em forma de gráficos:

- Para os resultados da avaliação do desempenho quanto ao espaço, solicita-se que, após a inclusão das n chaves, os valores obtidos de $\alpha^{\text{médio}}$ e p^* estejam no eixo y e os de $\alpha^{\max}$ e p estejam no eixo x. Serão gerados, portanto, dois gráficos do tipo linha. Notar que para o gráfico de $\alpha^{\text{médio}}$, médias obtidas para os resultados correspondentes a diferentes valores de p^* deverão ser consideradas, e vice-versa.

- Similarmente aos gráficos do item anterior, para avaliação do número de acessos, os valores de C e S devem estar no eixo y em função dos valores de $\alpha^{\max}$ e p.

- Por fim, os resultados de $\alpha^{\text{médio}}(i)$, $p^*(i)$ e $L^{\max}(i)$ deverão estar num único gráfico no eixo y, em três linhas distintas, em função do valore de i dispostos no eixo x.

O relatório deverá conter discussões sobre os resultados encontrados, demonstrando que houve entendimento das características observadas.

3 Observações

O trabalho pode ser feito em equipes, com no máximo três membros. Espera-se que até dia 25/09/2023 as equipes estejam definidas. Após esta data, equipes não poderão ser modificadas. Nomes não incluídos até esta data em alguma equipe serão interpretados como equipes individuais. A entrega do trabalho deverá conter código e relatório. Tanto as informações sobre as equipes quanto o trabalho deverão ser entregues através do Ava Moodle, na página disciplina, em local indicado, respeitando a data limite lá definida.

Referências

[1] Witold Litwin. Linear hashing: A new tool for file and table addressing. In *Very Large Data Bases Conference*, 1980.

[2] Alan Tharp. *File Organization and Processing.* John Wiley & Sons, Inc, 1988.