UNIVERSIDADE FEDERAL DA BAHIA

JEAN LOUI BERNARD
LUIZ GONZAGA SANTANA DOS SANTOS
IZAK ALVES GAMA

ESTUDO EXPERIMENTAL DO MÉTODO HASHING LINEAR

SALVADOR - BA
2023

JEAN LOUI BERNARD
LUIZ GONZAGA SANTANA DOS SANTOS
IZAK ALVES GAMA

ESTUDO EXPERIMENTAL DO MÉTODO HASHING LINEAR

Trabalho apresentado ao curso de Estrutura de Dados e Algoritmos II como requisito para obtenção de nota.

Docente: Dr. George Marconi de Araujo Lima

SALVADOR - BA
2023

SUMÁRIO

1. Resumo e métodos

Hashing Linear (HL) é uma estrutura de dados descrita pela primeira vez por Litwin [1] em 1980. Ela consiste na implementação de uma tabela *hashing*, a qual pode aumentar ou diminuir seu espaço alocado dinamicamente à medida que chaves são inseridas ou removidas, respectivamente, evitando o custo de reorganizá-las periodicamente. Seu funcionamento se baseia em múltiplas funções *hashing* que mapeiam as chaves nas listas de páginas, que por sua vez são alocadas linearmente. O número de páginas por lista cresce/diminui proporcionalmente ao número de colisões. Desse modo, este trabalho seguirá o método proposto por Litwin para a implementação de um algoritmo de HL, mantendo, porém, o mesmo tamanho das páginas de dados para as páginas de *overflow*.

Ademais, o relatório tem como objetivo avaliar o desempenho do método HL em 3 casos: espaço, número médio de acessos e inclusão. A avaliação será feita a partir da análise do seu comportamento em função do tamanho da página (p) e do fator de carga máximo (α^{max}), onde tais parâmetros serão variados durante a análise, com $p \in \{1,\ 5,\ 10,\ 20,\ 50\}$ e $\alpha^{max} \in \{0,2;\ 0,3;\ \cdots;\ 0,9\}$. Serão consideradas $n = 1000 \cdot p$ chaves aleatórias por experimento, que por sua vez serão repetidos 10 vezes. Os resultados para cada valor n será a média das 10 repetições.

2. Experimentos

Para os experimentos, implementou-se um algoritmo de HL na linguagem de programação C (presente junto com este documento). Além disso, utilizou-se algoritmos para monitorar o comportamento do HL, escritos em C e *Python*, a fim de obter os dados necessários para a realização dos experimentos descritos a seguir:

2.1. Desempenho quanto ao espaço

Este experimento tem o objetivo de avaliar a atitude do método HL quanto à alocação de espaço para inserção de chaves, na qual observamos o comportamento do

$\alpha^{médio}$ e da média de páginas por lista (p^*), em função do parâmetro α^{max}. Para tal, utilizamos o algoritmo em C para inserir n chaves aleatórias e calcular os parâmetros com base no uso e estado final do hashing em cada teste — fazendo a média dos resultados entre 10 repetições do teste — onde estes são obtidos da seguinte forma:

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

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

2.2. Desempenho quanto ao número médio de acessos

Neste experimento, procurou-se avaliar o número médio de acessos para recuperar $k = [0, 2n]$ chaves, considerando busca com (C) e sem (S) sucesso. A fim de calcular C, foram escolhidas aleatoriamente k chaves das n que foram incluídas. Já para S, considerou-se novamente k chaves aleatórias, porém distintas das n incluídas. Logo, C e S podem ser definidos formalmente do seguinte modo:

Sejam K^C e K^S os conjuntos de chaves para o cálculo de C e S, respecitavente.

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

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

Dessa forma, como descrito na seção 1, aferiu-se os resultados de C e S em função do tamanho da página e do fator de carga máximo, Então, os resultados foram dispostos em 2 gráficos tipo linha em função dos valores de p e α^{max}, respectivamente, considerando a média das aferições de C e S obtidas com a variação do outro parâmetro a fim de conferir legibilidade e representatividade aos dados. Ainda, um outro gráfico foi produzido para o número médio de acessos em função do tamanho da página, porém, desta vez com curvas de C e S para alguns α^{max} fixos.

2.3. Desempenho quanto à inclusão de registros

O objetivo deste experimento é analisar o comportamento do HL à medida que o número de registros cresce. Diferentemente dos experimentos descritos anteriormente, os valores de p e α_{max} não serão variáveis, e sim fixos em $p = 10$ e $\alpha_{max} = 0,85$, como sugerido nas especificações do trabalho. Com isso, temos que o número de registros a ser inseridos (n) variará de 1 à 10000. Cada item inserido é um número inteiro aleatório diferente dos já inseridos anteriormente, na qual pertence ao intervalo [0, 500000).

No experimento, serão computados os seguintes dados para o i-ésimo elemento inserido, com $i \in \{1,\ 2,\ \cdots,\ n\}$:

- $\alpha^{médio}(i) = \frac{\Sigma\ espaço\ ocupado\ em\ cada\ página}{espaço\ total\ alocado}$;
- $p^{*}(i) = \frac{\Sigma\ número\ de\ páginas\ em\ cada\ lista}{número\ total\ de\ listas}$ (média de páginas por lista);
- $L^{max}(i) = número\ de\ páginas\ na\ maior\ lista$;

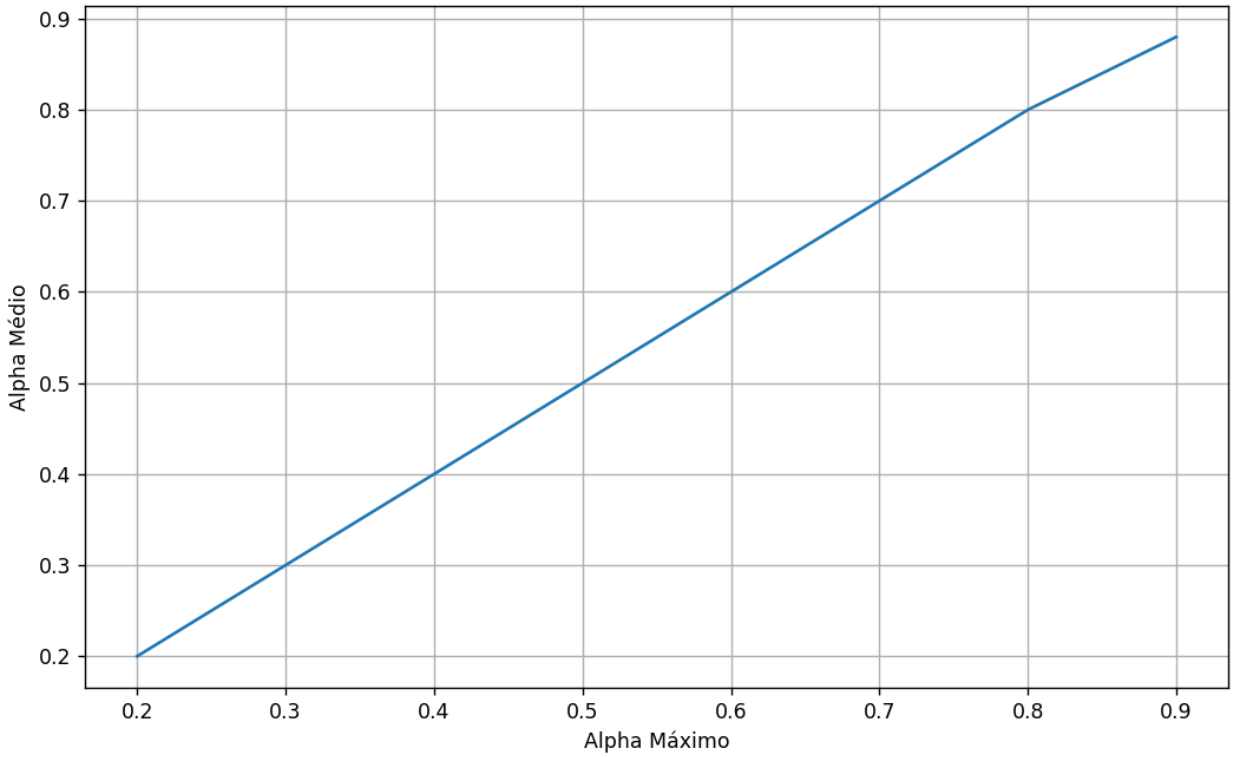
Assim como especificado na seção 1, o experimento será repetido 10 vezes e os resultados serão a média dos valores encontrados para cada i. Os resultados de $p^{*}(i)$ e $L^{max}(i)$ em função da quantidade de registros serão representados em um único gráfico, enquanto os resultados de α^{max} em gráfico próprio, já que esse último varia dentro do intervalo $(0;\ 0,85)$, enquanto os demais variam em intervalos bem maiores.

3. Resultados e análises

3.1. Análise de desempenho quanto ao espaço

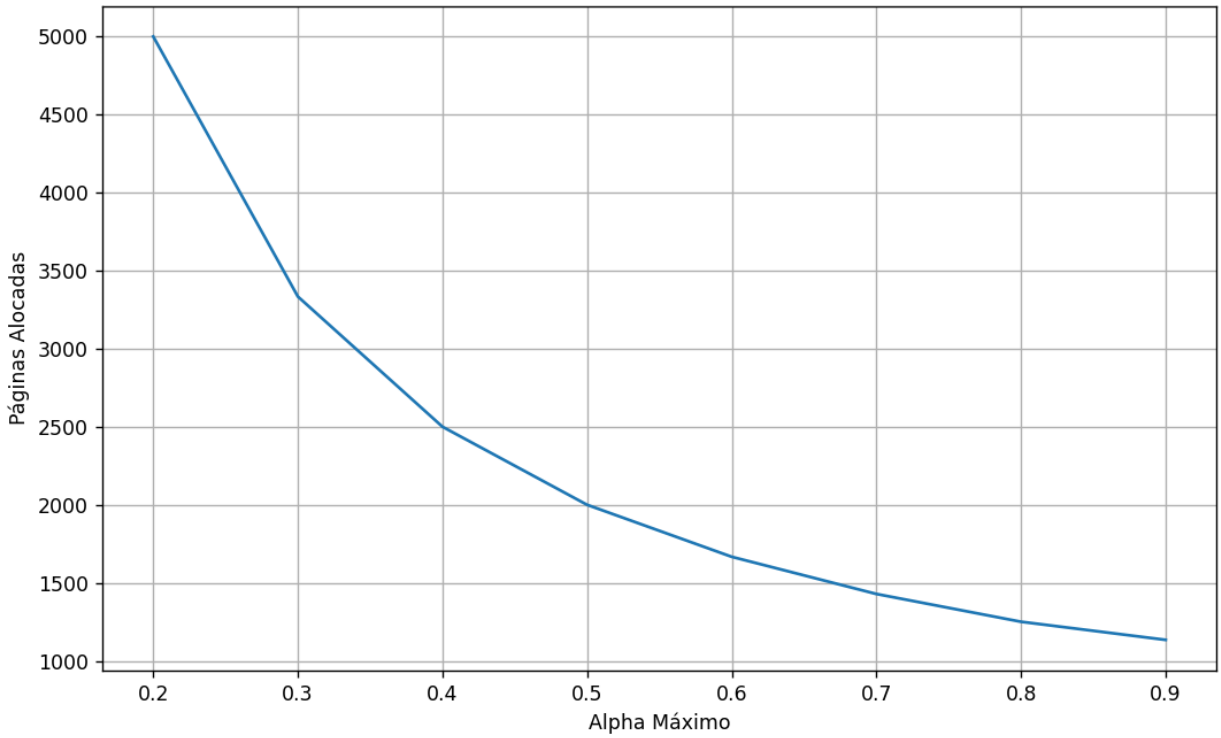
Seguem abaixo os gráficos do desempenho do HL em relação ao espaço alocado, onde o primeiro e o segundo gráfico mostra o comportamento do ($\alpha^{médio}$) e a quantidade média de páginas alocadas, e o terceiro gráfico mostra o comportamento do (p^{*}), todos eles em função do (α^{max}).

Figura 3.1.1 - $\alpha^{médio}$ em função de α^{max}



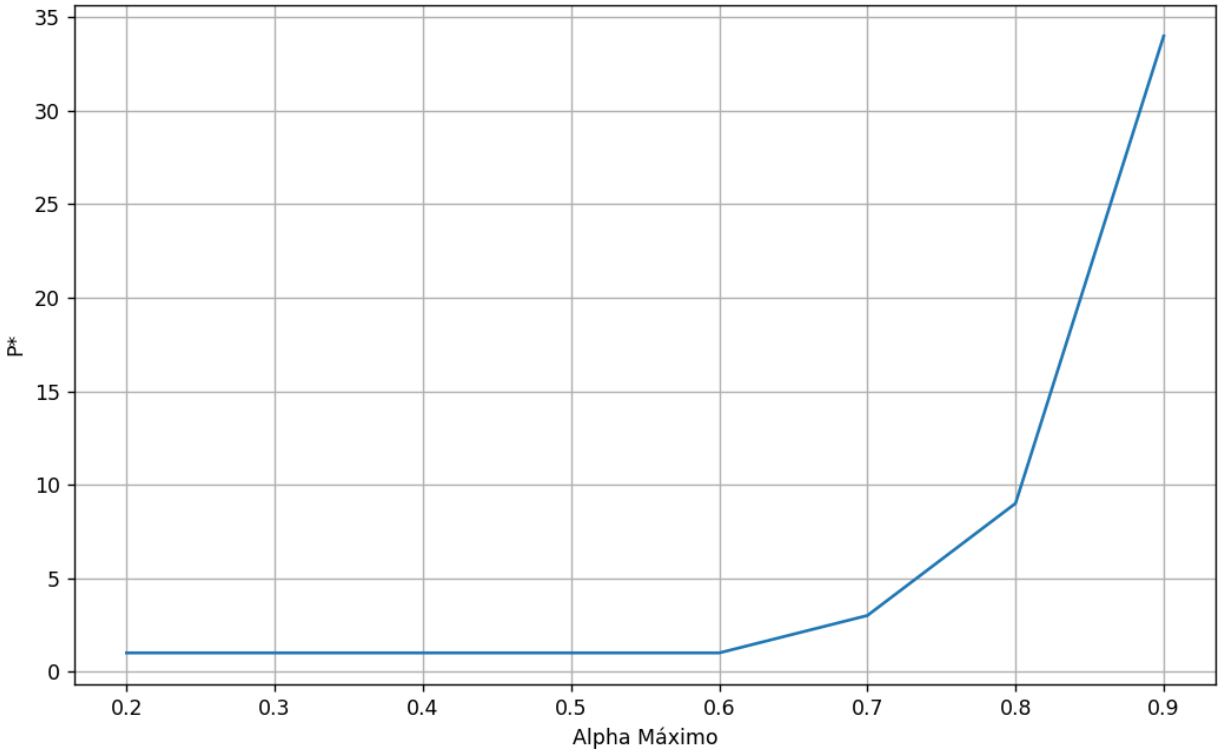
Fonte: Autoria própria, 2023.

Figura 3.1.2 - Quantidade de páginas alocadas em função do α^{max}



Fonte: Autoria própria, 2023.

Figura 3.1.3 - Média de páginas (p*) por lista em função do α^{max}



Fonte: Autoria própria, 2023.

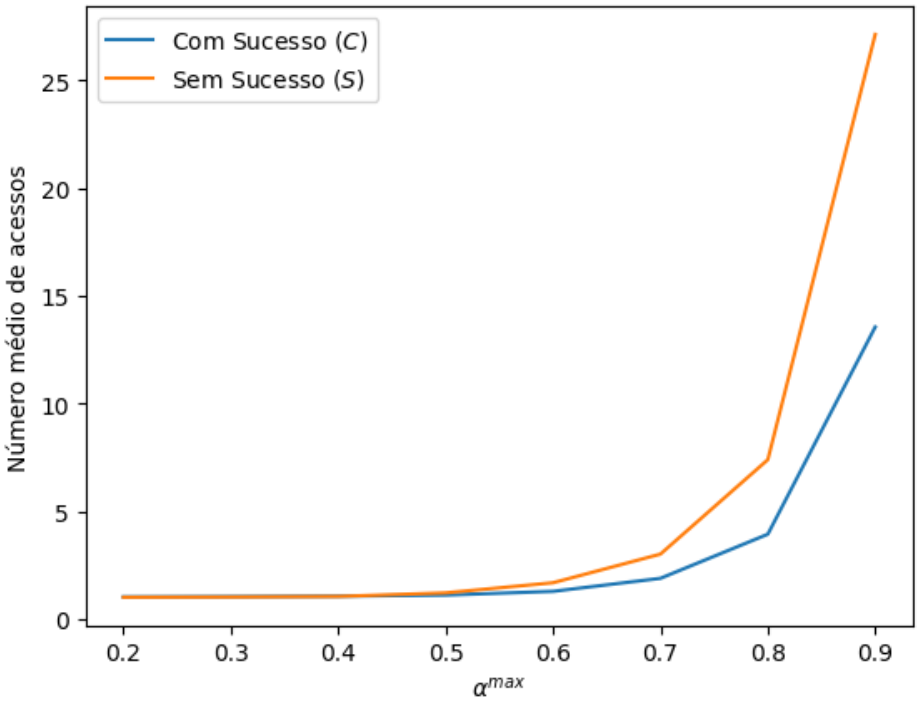
Em todos os experimentos descritos acima, obteve-se a média dos resultados para cada (α^{max}). Como foi observado na figura 3.1.1, o valor de ($\alpha^{médio}$) cresce com forte correlação linear em função do (α^{max}), o que nos indica que o (α^{max}) está diretamente associado ao consumo de memória excedente. Além disso, nas figuras 3.1.2 e 3.1.3, observa-se que, à medida que (α^{max}) aumenta, existe um decréscimo da quantidade de páginas alocadas e um crescimento do número médio de páginas por lista (p*), que cresce rapidamente a partir de ($\alpha^{max} = 0,7$).

Desse modo, como era esperado, podemos concluir que valores menores para o (α^{max}) acabam gerando um grande desperdício de espaço, porém uma grande distribuição das páginas no método *hashing*. Por outro lado, valores maiores para α^{max} geram menos desperdício de espaço, porém tendem a ter uma baixa distribuição das páginas, criando listas muito densas, o que pode afetar na busca e inserção de chaves.

3.2. Análise de desempenho quanto ao número médio de acessos

Primeiramente, avaliou-se o número médio de acessos para C e S à medida que α^{max} aumentava. Conforme demonstrado na figura 3.2.1, o número de acessos permanece aproximadamente constante para buscas com sucesso, até começar a crescer rapidamente a partir de $\alpha^{max} = 0,7$. De forma semelhante, essa tendência também ocorre para buscas sem sucesso, com a diferença do crescimento começar em $\alpha^{max} = 0,6$, além de ser significativamente mais acelerado.

Figura 3.2.1 - Número médio de acessos em função do α^{max}



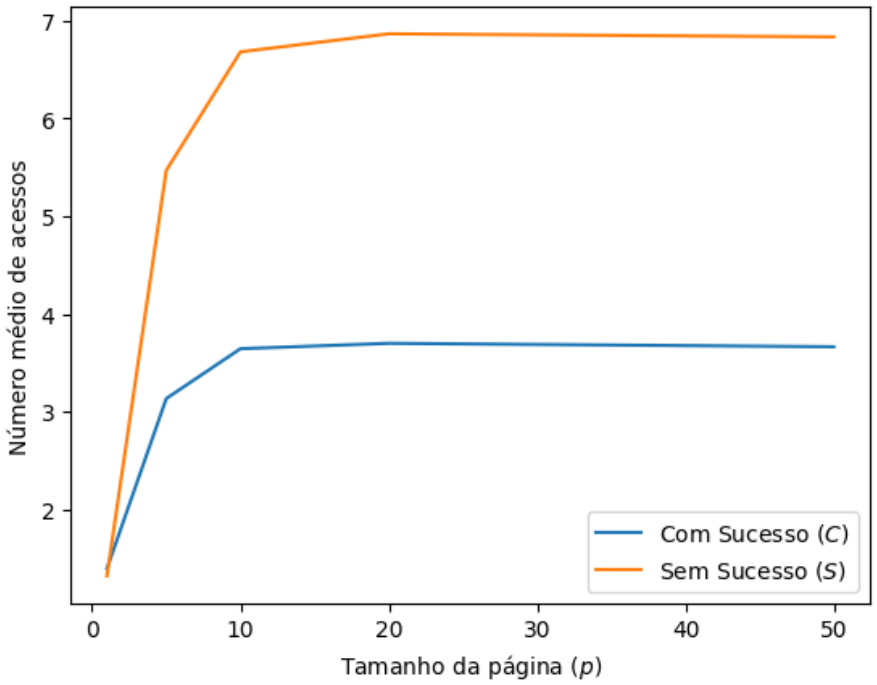
Fonte: Autoria própria, 2023.

De fato, esses resultados já eram esperados. Para $\alpha^{max} \to 1$, a criação de novas listas de páginas se torna menos frequente, o que resulta em um agrupamento das chaves em um número menor de listas e, consequentemente, um número maior de páginas em cada lista. Assim, para recuperar uma chave presente na estrutura, será

necessário, em média, um número maior de acessos. Além disso, se a busca não obtiver sucesso, então será preciso acessar todas as páginas de uma determinada lista, logo o número médio de acessos será ainda maior.

A segunda avaliação consistiu em estipular o número médio de acessos para C e S variando o tamanho da página (p). É possível notar pela figura 3.2.2 que C e S apresentam um valor menor para $p = 1$. Além disso, a curva de ambos apresenta um comportamento praticamente constante a partir de $p = 10$.

Figura 3.2.2 - Número médio de acessos em função do tamanho da página

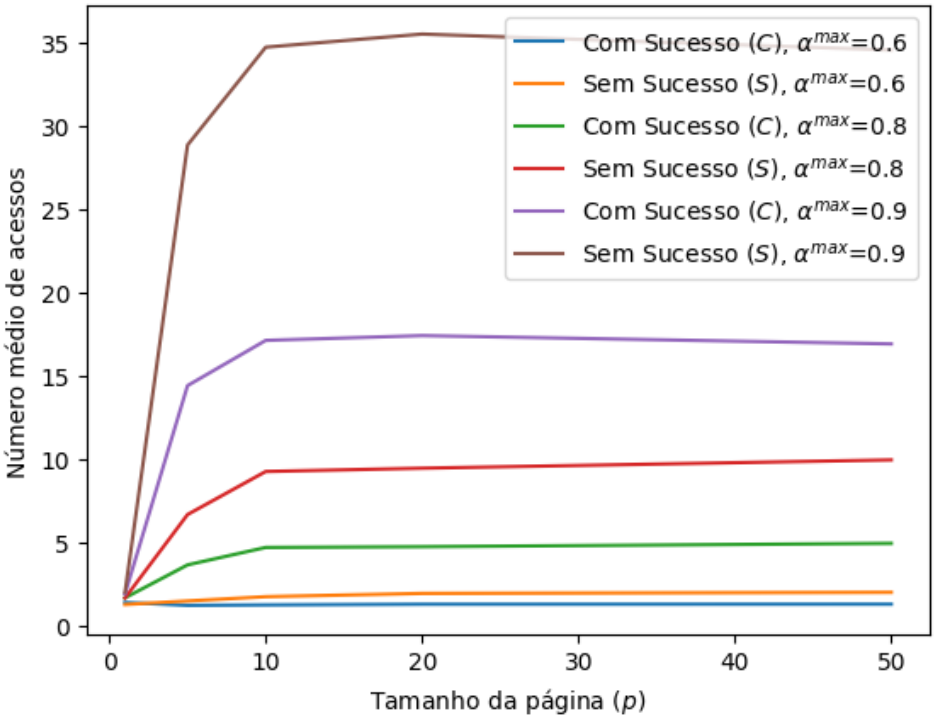


Fonte: Autoria própria, 2023.

De maneira simetricamente oposta à análise da variação de α^{max}, a razão de C e S terem o número médio de acessos inferior para $p = 1$ — e em menor grau para $p = 5$ — é a produção de mais listas de páginas durante as inserções devido ao rápido preenchimento de cada página. Isso tem como consequência a redução do número de páginas por lista, que por sua vez reduz o número médio de acessos. Ainda, pelo mesmo motivo da análise anterior, o número médio de acessos de S é consideravelmente maior que o de C para $p > 1$.

Logo, poderia-se esperar que à medida que p aumentasse, a produção de listas de páginas diminuísse e o número médio de acessos aumentasse. No entanto, a constância das curvas observada para $p \geq 10$ demonstra que, a partir desse ponto, o aumento proporcional a p do número de chaves inseridas compensou o aumento de p e, portanto, não provocou alterações significativas no número de listas. Desse modo, como mostra a figura 3.2.3, o tamanho da página não explica o comportamento do número médio de acessos para $p \geq 10$, sendo o α^{max} o parâmetro mais relevante para a sua explicação, considerando os valores de p e α^{max} adotados no experimento.

Figura 3.2.3 - Número médio de acessos em função do tamanho da página, com α^{max} fixo.

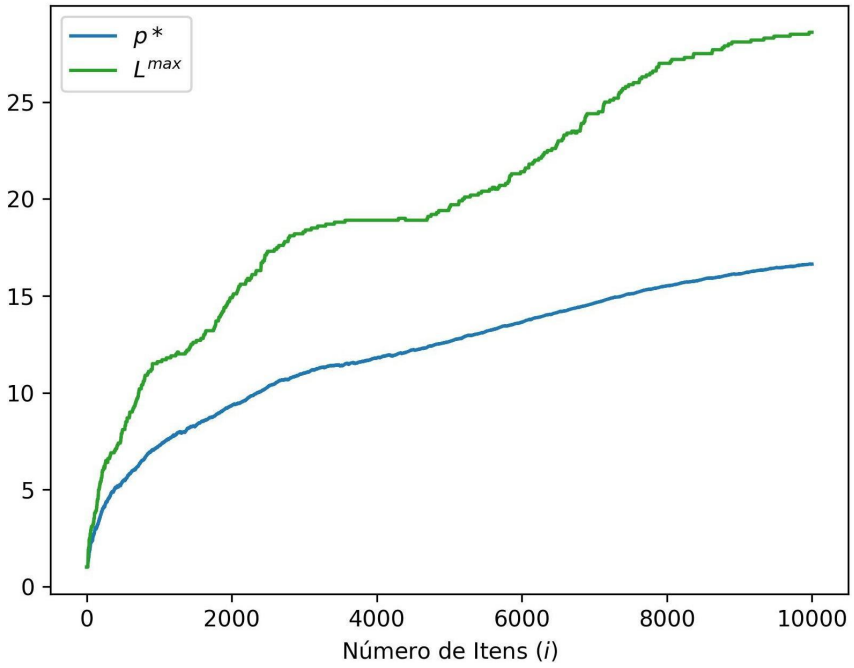


Fonte: Autoria própria, 2023.

3.3. Análise de desempenho quanto à inclusão de registros

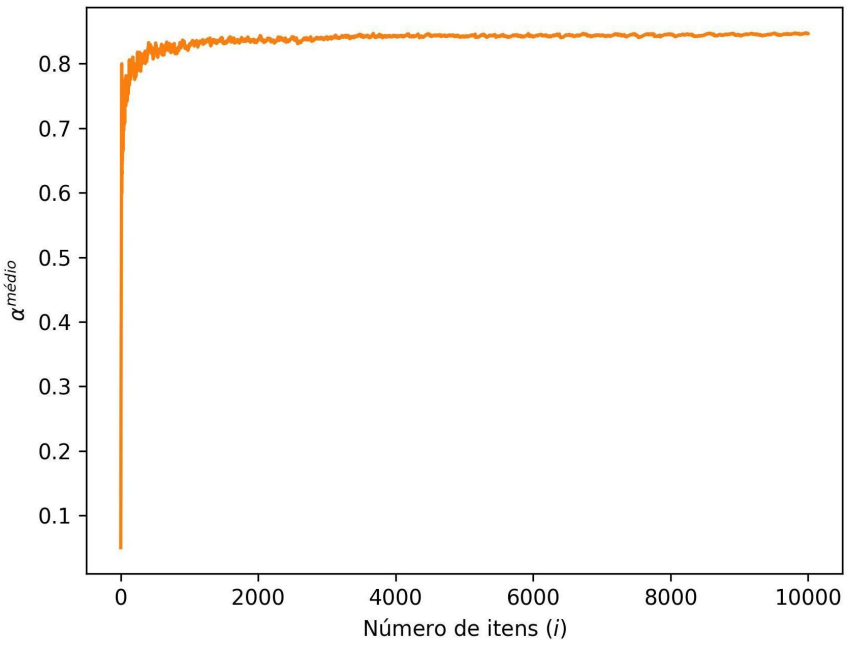
Seguem abaixo os gráficos do desempenho do HL em relação ao número de itens inseridos, onde o primeiro gráfico mostra o comportamento de p^* e L^{max} e o segundo o comportamento de $\alpha^{médio}$.

Figura 3.3.1 - p^* e L^{max} em função do número de itens inseridos



Fonte: Autoria própria, 2023.

Figura 3.3.2 - $\alpha^{médio}$ em função do número de itens inseridos



Fonte: Autoria própria, 2023.

Em relação a média de páginas por lista (p^*), percebe-se que é crescente de acordo com a quantidade de itens inseridos (i). Nota-se que p^* cresce lentamente, se comparado com i, pois mesmo com i = 10000, p^* em média é ligeiramente maior que 15, considerando páginas de tamanho 10. Quanto maior valor de i, maior é a diferença. Aparentemente, p^* cresce logaritmicamente em função de i. Para a inserção de um novo item, visto que a estrutura de dados já possui 10000 registros, em média colidiria com um pouco mais do que outros 150 registros, ou seja, em média uma inserção colide com um pouco mais que 1,5% dos itens armazenados, o que é um número bem baixo.

Já em relação ao número de páginas na maior lista (L^{max}), percebe-se que também é crescente de acordo com i, além de crescer lentamente, pois em i = 10000, L^{max} é ligeiramente menor que 30. Aparentemente L^{max} também cresce logaritmicamente. Comparando com p^*, percebe-se que a diferença entre ambos também é crescente, visto que L^{max} cresce mais rápido. Para a inserção de um novo

item, também considerando a existência 10000 registros, colide-se com no máximo um pouco menos que outros 300 registros (em média), ou seja, uma inserção colide com no máximo um pouco menos que 3% dos itens armazenados (em média), o que também é um número baixo.

Em relação ao fator de carga ($\alpha^{médio}$), percebe-se que na maior parte do processo de inserção permanece variando ligeiramente abaixo do fator de carga máximo (α^{max}), que é 0,85. É possível notar que quando há poucos itens inseridos, $\alpha^{médio}$difere-se um pouco mais de α^{max}, havendo um crescimento gradual, porém rápido, com grande variação de $\alpha^{médio}$. Já quando há muitos itens inseridos, o valor de $\alpha^{médio}$ aproxima-se bastante de α^{max} e varia com menos frequência. Isso ocorre devido ao tamanho da página utilizada ser relativamente pequeno.

Quando há poucos itens inseridos, adicionar espaço para mais 10 registros é relativamente significativo em relação ao espaço total alocado, então os valores de $\alpha^{médio}$ antes e depois da adição da página podem até serem próximos, mas há uma diferença considerável. Por isso $\alpha^{médio}$ varia bastante (vide a fórmula de $\alpha^{médio}$ na seção 2.3). Já quando há muitos itens inseridos, a adição de espaço para mais 10 registros não é relativamente significativa, pois é um valor pequeno em relação à quantidade de registros, então os valores de $\alpha^{médio}$ antes e depois da adição da página são muito próximos, não havendo uma diferença suficientemente considerável. Por isso $\alpha^{médio}$ costuma variar pouco, assemelhando-se à uma reta.

4. Conclusão

Neste relatório foram examinados aspectos de performance importantes do método *hashing* linear. Na análise do desempenho quanto ao espaço, a partir dos resultados atingidos, concluiu-se que quanto maior o valor de α^{max}, maior o aproveitamento da memória alocada. Em contrapartida, a estrutura tende a ter mais páginas em cada lista.

Em decorrência desse fato, na análise do desempenho quanto ao número médio de acessos, observou-se que quanto maior α^{max}, maior o número médio de acessos para buscas com e sem sucesso. Ademais, constatou-se que o tamanho da página, quando maior ou igual a 10, não apresenta correlação com o comportamento da curva de C e S.

Em consonância com os resultados anteriores, na análise do desempenho durante a inclusão de n chaves, aferiu-se que o tamanho da maior lista, assim como a média de páginas por lista, cresce com o aumento do número de registros inseridos. Ainda, notou-se que o aproveitamento de espaço permanece muito próximo ao α^{max} adotado quando são inseridos mais de 1000 itens, o que resulta em um bom aproveitamento de espaço.

Assim, a partir dos experimentos realizados, foi possível averiguar que os parâmetros que produzem, ao mesmo tempo, um grande aproveitamento do espaço e um baixo número de acessos é $\alpha^{max} = 0,7$ e $p \geq 10$ para um número de registros de até 10000. Portanto, este relatório logrou os objetivos propostos, proporcionando uma maior compreensão das propriedades mais influentes do método HL.

5. Referências

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