Implementação de uma Tabela Hash para Indexação de Documentos

João Vitor Coelho Oliveira
João Pedro dos Santos Ferraz

30 de março de 2025

Universidade Federal de Ouro Preto
Ciência da Computação
Estruturas de Dados I

Resumo

Este relatório descreve a implementação de uma Tabela Hash para indexação de palavras em documentos utilizando a linguagem C. O trabalho apresenta os conceitos utilizados, detalhes da implementação, metodologia de validação e os resultados obtidos.

1 Introdução

A indexação de palavras em documentos é uma técnica fundamental em busca de informação. Neste trabalho, implementamos uma Tabela Hash para armazenar palavras extraídas de documentos, associando-as aos respectivos nomes de arquivos nos quais aparecem.

2 Considerações Iniciais

- Ambiente de desenvolvimento: GCC e Valgrind.
- Linguagem utilizada: C.
- Documentação gerada em Overleaf (LaTeX).

3 Metodologia

A implementação foi dividida nas seguintes partes principais:

- Criação da estrutura da Tabela Hash.
- Funções para inserção e busca de palavras na tabela.
- Estruturação de um vetor para armazenar a ordem de inserção das palavras.
- Impressão dos resultados de acordo com a ordem correta.

3.1 Estrutura do Código

O código implementa uma estrutura `HashTable` que armazena palavras e seus respectivos documentos. A tabela utiliza uma função de dispersão para distribuir as palavras de maneira eficiente.

3.1.1 Funções Implementadas

- `criaHashTable`: Inicializa a tabela hash.
- `h`: Calcula a posição de armazenamento de uma palavra.
- `inserePalavra`: Adiciona uma palavra na tabela, associando-a a um documento.
- `imprimeHash`: Exibe as palavras e seus documentos na ordem correta de inserção.
- `buscaPalavra`: Procura uma palavra na tabela hash e retorna os documentos associados.
- `removePalavra`: Remove uma palavra da tabela hash, se presente.
- `liberaHashTable`: Libera a memória alocada para a tabela hash.
- `contaPalavras`: Conta o número total de palavras armazenadas na tabela.

3.1.2 Principal Função e Seu Funcionamento

A principal função do código é a `inserePalavra`, responsável por adicionar novas palavras na Tabela Hash, garantindo que elas sejam indexadas corretamente. O funcionamento ocorre da seguinte forma:

- O nome do documento e a palavra são recebidos como entrada.
- A `hashFunction` é chamada para calcular o índice apropriado na tabela.
- Caso a palavra já exista na tabela, o documento é adicionado à sua lista de ocorrências.
- Se a palavra não existir, uma nova entrada é criada na tabela com a palavra e seu documento correspondente.
- A inserção é feita de forma eficiente para minimizar colisões e garantir a rapidez na busca.

Esta função é essencial para a indexação correta dos documentos, garantindo que cada palavra seja associada aos arquivos onde aparece.

4 Lógica de Resolução do Problema

O programa recebe palavras e seus respectivos documentos como entrada e as armazena na Tabela Hash. A função `inserePalavra` verifica se a palavra já existe na tabela antes de adicioná-la. A função `imprimeHash` exibe as palavras na ordem correta.

5 Resultados Obtidos

Os testes foram realizados utilizando conjuntos de palavras de diferentes tamanhos. Os resultados confirmaram que a Tabela Hash armazena e recupera corretamente as palavras e seus documentos. Além disso, foi utilizado a ferramenta Valgrind para monitorar o tempo de execução e a utilização de memória.

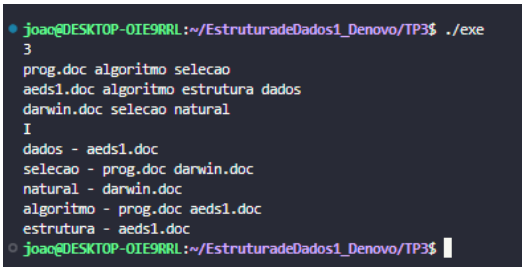

Figura 1: Exemplo de Saída da Tabela Hash

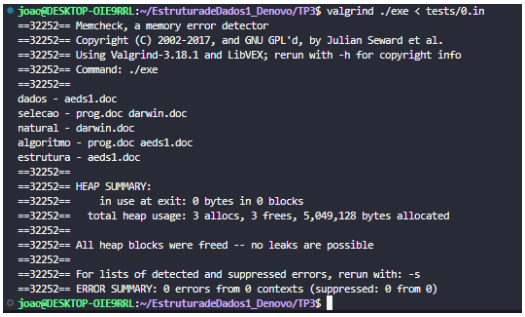

Figura 2: Uso do Valgrind para verificação

5.1 Casos de Teste

Foram utilizados diferentes conjuntos de palavras e arquivos para validar o funcionamento da Tabela Hash. Todos os testes retornaram os resultados esperados.

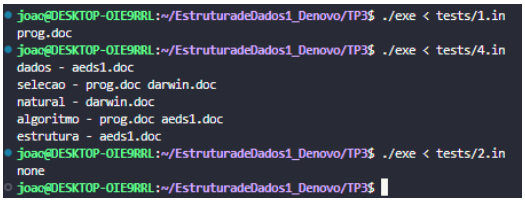

Figura 3: Casos de teste

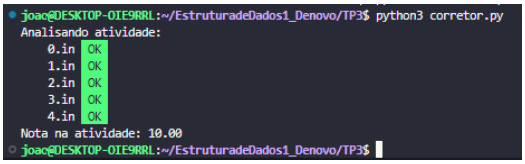

Figura 4: utilização do corretor.py

6 Conclusão

A implementação da Tabela Hash demonstrou ser eficiente para indexação de palavras. Alguns desafios encontrados incluem:

- Lida com colisões na tabela hash.
- Impressão das palavras na ordem correta.
- Gerenciamento da memória dinâmica.

7 Referências

- Hash Tables. Disponível em: `https://en.wikipedia.org/wiki/Hash_table`.
- Valgrind. Disponível em: `https://valgrind.org/`.
- Overleaf. Disponível em: `https://www.overleaf.com/`.