Instituto Federal Sudeste de Minas Gerais – Campus Barbacena
Curso de Tecnologia em Sistemas para Internet
Disciplina: Estruturas de Dados II
Prof. Wender Magno Cota

Primeiro Trabalho Prático I – Primeira Parte

Construção de Índice para uma Coleção de Documentos

O aumento no volume de dados armazenados na forma de textos tem acompanhado o rápido crescimento das bibliotecas digitais modernas, como enciclopédias e a própria World Wide Web. Para se recuperar de forma eficiente a informação armazenada em grandes bases de dados textuais, a única solução prática é construir estruturas auxiliares, denominadas índices, de forma a tornar computacionalmente viável a utilização de sistemas de recuperação de informação. Existem várias técnicas de indexação, entre elas Árvore PATRICIA, PAT Array, Arquivos Invertidos, etc. O arquivo invertido é a principal estrutura de dados utilizado para construir um índice para acelerar o processamento de consultas em máquinas de busca. Um arquivo invertido é um mecanismo orientado por palavras para indexar uma coleção de texto com o objetivo de aumentar a velocidade de pesquisa. A estrutura do arquivo invertido é composta de duas partes: (i) o vocabulário e (ii) as listas invertidas. O vocabulário é o conjunto de todas as palavras distintas contidas no texto da coleção. Para cada uma destas palavras, uma lista de todos os documentos onde a palavra ocorre, juntamente com a frequência da palavra em cada documento é armazenada. Essas listas são chamadas de listas invertidas. A seguir será mostrado um arquivo invertido construído a partir de uma coleção contendo 2 documentos.

arquivo1.txt
Quem casa quer casa. Porem ninguem casa.
Ninguem quer casa tambem. Quer apartamento.

arquivo2.txt
Ninguem em casa. Todos sairam. Todos.
Quer entrar? Quem? Quem?

Vocabulário	**Lista Invertida**	
apartamento	1 1	
casa	4 1	1 2
em	1 2	
entrar	1 2	
ninguem	2 1	1 2
porem	1 1	
quem	1 1	2 2
quer	3 1	1 2
sairam	1 2	
tambem	1 1	
todos	2 2	

Projete um sistema para produzir um índice invertido. Tal sistema deve processar uma coleção de documentos existentes em uma determinada pasta e gerar o arquivo invertido. Associe a cada documento um **doc_id** único e associar, em memória, este identificador com o nome do documento.

Nota:

a) Uma palavra é considerada como uma sequência de letras e dígitos, começando com uma letra. Portanto, ignore sinais de pontuação. Você pode assumir que os textos não terão acentuação.

b) Palavras com letras em maiúsculas devem ser primeiramente transformadas para minúsculas antes da inserção no índice.

c) Apenas os primeiros N caracteres devem ser retidos nas chaves. Assim, duas palavras que não diferem nos primeiros N caracteres são consideradas idênticas.

d) Uma palavra pode ocorrer múltiplas vezes na mesma linha de um documento, ou mesmo em múltiplas linhas de um mesmo documento.

A construção do índice deve ser implementado usando as estruturas abaixo:

1) hash com endereçamento aberto (linear);
2) hash com resolução de conflito via árvore binária sem balanceamento;
3) Árvore AVL;
4) Árvore B.

Realizar medidas o desempenho, em termos do tempo de execução e do consumo de memória, de cada um dos TADs para diferentes cenários. Utilize contadores em como chamadas do sistema para medições das estatísticas solicitadas.

Obs.

a) Data de Entrega: 20/04/2012;
b) Valor: 10,0 pontos.