Trabalho Prático

Programação e Desenvolvimento de software II – 2019/2

Martinelle Araujo dos Santos
Alexandre Nascimento Junior

23 de Novembro de 2019

1 Introdução

O método de conversão de dados é uma técnica de transferência dados para um formato que seja aceitável, o processo fornece uma excelente oportunidade de avaliar as suas informações atuais e de tomar as melhores decisões a respeito. A Máquina de Busca(MB) é um algorítimo que readquire informações em uma base de dados. A MB recebe como entrada uma expressão de busca, processa a mesma e dá como saída os documentos mais revelantes para a consulta fornecida. Para que haja uma maior eficiência, utiliza-se tipo abstatros de dados para que ocorra a melhor eficiência no processo de recuperação.

Este trabalho tem por objetivo a implementação de um índice invertido. O algorítimo proposto é uma estrutura de dados que mapeia termos às suas ocorrências em um documento ou conjunto de documentos, e armazenados em um banco de dados. É uma estratégia de indexação que permite a realização de buscas precisas e rápidas, em troca de maior dificuldade no ato de inserção e atualização de documentos.

2 Implementação

Abaixo segue-se uma breve explicação em alto nível sobre como o algoritmo foi implementado.

Entrada: Para a implementação do algorítimo devemos iniciar com a entrada dos arquivos. Deve-se entrar com **o nome de cada arquivo (incluindo a extensão) separados por um espaço**.

Leitura: Com a entrada dos arquivos iniciamos a leitura e o tratamento das palavras. O tratamento é feito para que existam apenas letras de a-z e números de 0-9. De modo que, transformamos as letras maiúsculos em minusculas e retiramos qualquer carácter especial, exceto no caso do hífen que é transformado em espaço.

Frequência, Importância e Incidência: A medida que cada palavra passa pelo tratamento, ela já recebe um flag de que pertence ao arquivo que está sendo lido no momento. Com isso podemos calcular em quantos arquivos a palavra aparece ao final da leitura dos arquivos, apenas somando os valores de um vetor em que cada posição representa um arquivo e a posição recebe "1"se a palavra está naquele arquivo.

Também simultaneamente a leitura do arquivo são incrementadas todas as vezes em que a palavra aparece naquele arquivo em questão. Ao final da leitura, tendo essas informações, o calculo da importância, como especificado na documentação, foi facilitado. Nessa etapa, o uso de maps e vectors foi essencial.

Índice Invertido: No índice invertido usamos novamente o map com chave **palavra** e valor**vetor**, este vetor é mesmo citado anteriormente, contendo "1"nas posições correspondentes aos arquivos aos quais a palavra está presente.

Coordenadas: Para cada arquivo recebido na entrada, calculamos um vetor para representa-lo. Aqui, cada coordenada do vetor diz respeito a uma palavra.

Consulta: Quando uma palavra for consultado ela sera tratada da mesma maneira que os arquivos inicias, salvando as palavras em um vetor e fazendo importância vezes frequência teremos a coordenada do novo vetor das palavra pesquisadas. Em seguida é ordenar e comparar o vetor com os vetores do arquivos já existentes.

3 Conclusão

Como conclusão após os testes observa-se o esperado, o índice invertido, é portanto um algorítimo que tem um desempenho rápido e eficiente. Uma busca normal exigira percorrer cada documentos a procura da palavra pesquisa, enquanto que, com o uso do índice invertido, você pode ir de formar eficaz no documento que você busca.

O uso do índice tem como objetivo deixar as buscas mais eficaz mas vem como um trabalho manual adicional, pois precisa de sempre esta em completa manutenção desta lista, já que, precisasse sempre mantela atualizada conforme novos documentos são inseridos, alterados ou excluídos.