Instituto de Computação (ICOMP)
Universidade Federal do Amazonas (UFAM)

Disciplina: Recuperação de Informação
Discente: Maria Ivanilse Calderon Ribeiro
E-mail: {ivanilse.calderon@icomp.ufam.edu.br}
Matricula: 3150047

Trabalho Prático: Implementação de máquina de busca para uma coleção de documentos Utilizando o Modelo Vetorial

Resumo: *Este relatório apresenta implementação de uma máquina de busca utilizando o Modelo Vetorial, bem como a análise dos resultados apresentados utilizando as métricas de avaliação MAP e P@10.Tal implementação e análise é proposta na disciplina Recuperação de Informação do Programa de Pós-Graduação em Informática da Universidade Federal do Amazonas – UFAM.*

1. Objetivos

O trabalho consiste no trabalho prático da disciplina, onde é proposta a implementação de uma máquina de busca utilizando o Modelo Vetorial para busca na coleção de documentos CFC. A base para testes está dividida em 6 arquivos contendo 1.239 documentos publicados entre 1972 e 1974, um arquivo com 100 consultas e seus respectivos relevantes. Em se tratando das avaliações de relevância é importante observar que estas foram feitas por 4 pesquisadores diferentes. Assim, é identificado 4 avaliações diferentes para cada par artigo/consulta existente.

O objetivo do trabalho é a implementar o Modelo Vetorial na Linguagem C++, a apresentar uma comparação dos resultados retornados utilizando as métricas MAP e P@10. Para tal, foram considerados como relevantes todos os documentos relacionados no arquivo de consultas, não levando em consideração a nota de relevância consideradas pelos avaliadores.

2. Tecnologias utilizadas

Para a implementação do código para a máquina de busca, bem como a realização dos experimentos foram utilizadas as seguintes especificações: Sistema Operacional Linux, distribuição Ubuntu (14.04), Linguagem de programação C++, Compilador GCC, Bibliotecas da linguagem, Notebook de arquitetura 64 bits, memória RAM 8 Gigas, processador core i7 - 4500U/ CPU @ 1.80GHz.

3. Implementação

A linguagem utilizada para a implementação do modelo foi C++. Foram utilizadas as bibliotecas e recursos que a linguagem dispõem. As principais estruturas de dados utilizadas nesse trabalho foram:

Map: estrutura utilizada, vez que na referida implementação é necessário classificar e identificar exclusivamente os elementos, enquanto que os valores

mapeados armazenar o conteúdo associado a esta chave, sendo possível assim quantificar os pesos e frequências dos termos na coleção de documentos. Tal tipo de dado permite armazenar os termos formados por uma combinação de um valor de chave e um valor mapeado, seguindo uma ordem específica, permitindo assim a criação do índice invertido e da lista invertida. Sendo map<int, map<string, float> > matriz_peso; map<string, Consultas> hash_consulta; map<int, float> norma_colecao; algumas das estruturas implementadas na máquina de busca.

Vector: tendo a necessidade de uma estrutura que cresce dinamicamente, onde os elementos podem serem acessados por meio de um índice. O tipo foi utilizado para armazenar os documentos e as palavras limpas. Sendo vector<int> ranking; vector<int>::iterator it; vector<double>::iterator j; que permite manipulação do ranking, bem como a aplicação das métricas. A utilização de **iterator** foi importante, pois permitiu percorrer facilmente uma estrutura da STL de um extremo a outro. O iterator lembra um pouco a noção de ponteiro, mas não é um endereço.

4. Link da implementação no GitHub

https://github.com/Ivanilse/tp_ri.

5. Compilação e execução da implementação

Para que a execução do programa é necessário copiar os arquivo do programa, bem como os arquivo de stopwords e a base utilizada nas buscas. Esse estão organizados da seguinte forma:

a) No diretório **cfc** constam os arquivos: "cf74 ao cf79" que são os 6 arquivos com a base da coleção. O "cfquery" que é o arquivo da consulta. O arquivo com as stopwords "stopwords".

b) No diretório **src** constam os arquivos da implementação da máquina de busca, que são: "Consulta.cpp, Indice.cpp, Main.cpp"; "hash.txt, saida.txt, out.txt" e "Consulta.h; includes.h; Indice.h".

Por meio da utilização do terminal Linux, execute o compilador GCC e execute os comandos a seguir para execução e testes da máquina de busca:

g++ -std=c++0x -I. *.cpp -o out
./out > saida.txt

6. Experimentos realizados

Após realizar dos experimentos com a base disponibilizada, observa-se que:

a) O tempo de execução para as consultas apresenta-se aproximadamente em torno de 6000 milisegundos;

b) A média de avaliação para a métrica **Map** é de 27,74%;

c) A média de avaliação para a métrica **P@10** é de 42,20%.