CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
PROJETO E ANÁLISE DE ALGORITMOS

Última alteração: 4 de maio de 2006

Professor: Nivio Ziviani

Monitor: Fabiano C. Botelho

3º Trabalho Prático - 04/05/06 - 10 pontos

Data de Entrega: 26/05/06

Penalização por Atrazo: 1 ponto até 02/06/06 mais 1 ponto por dia útil a seguir

Observação: Toda a documentação deverá ser apresentada como uma página acessível via Web (apresente o link para acesso à documentação).

Casamento de Padrão

O objetivo deste trabalho é realizar experimentos em um conjunto de programas para recuperar ocorrências de padrões em arquivos constituídos de documentos comprimidos e não comprimidos. O sistema de programas recebe do usuário um texto contendo n caracteres, o padrão contendo m caracteres, o número de erros ($0 \leq k < m$), e imprime todas as ocorrências do padrão no texto.

Busca Aproximada em Arquivos Não Comprimidos

Nesta parte do trabalho você deverá utilizar os seguintes algoritmos:

- Algoritmo programação dinâmica para casamento aproximado.
- Algoritmo Shift-And para casamento aproximado.

O que fazer

Compare o desempenho dos dois algoritmos para valores de $k = 0, 1, 2, 3$:

- Medindo o número de comparações entre caracteres.
- Medindo o tempo de relógio.

Busca Exata em Arquivos Comprimidos

Nesta parte do trabalho você deverá comparar o desempenho do algoritmo de Boyer-Moore-Horspool (BMH) para para arquivos comprimidos e não comprimidos. Para comprimir o arquivo você deve utilizar o código de Huffman com marcação descrito em Ziviani (2004, Seção 8.2). Um arquivo comprimido com esse algoritmo permite que o padrão de busca seja comprimido e buscado diretamente no arquivo comprimido.

O que fazer

Compare o desempenho do algoritmo BMH para arquivos comprimidos e não comprimidos:

- Medindo o número de comparações entre caracteres.
- Medindo o tempo de relógio.

Como fazer

1. Utilize os arquivos de documentos da coleção TREC disponível em `/export/texto2/wsj`, com as seguintes características:

Coleção	Tamanho (Mb)
wsj88	109
wsj88_20	20
wsj89	2.8

   Um exemplo de registro deste arquivo é:

```
<DOC>
<DOCNO> WSJ890802-0125 </DOCNO>
<DD> = 890802 </DD>
<AN> 890802-0125. </AN>
<HL> Inside Track:
@  NCNB Director Sold Big Holding in July
@  Worth $7.4 Million More 4 Weeks Later
@  ----
@  By Alexandra Peers and John R. Dorfman
@  Staff Reporters of The Wall Street Journal </HL>
<DD> 08/02/89 </DD>
<SO> WALL STREET JOURNAL (J) </SO>
<CO> NCB BPCO TRN EGGS LABOR </CO>
<IN> STOCK MARKET, OFFERINGS (STK)
SECURITIES INDUSTRY (SCR) </IN>
<TEXT>
   Even insiders make mistakes.
Sometimes, big, fat, $7 million-dollar mistakes. ...
</TEXT>
</DOC>
```

2. Crie uma interface de uso do seu sistema parecida com a do sistema *agrep* disponível na rede do DCC.

3. A saída do programa deve mostrar a linha do documento que contém o padrão.

4. Utilize as consultas listadas no arquivo `/export/texto2/wsj/wsj89_consultas` para testar seu programa você deve realizar e mostrar o resultado obtido.

Observação: O sistema *agrep* disponível na rede do DCC pode ser um bom parâmetro para comparar o desempenho do programa BMH em relação a uma implementação considerada rápida.

O que deve ser entregue:

- Listagem dos programas implementados. Explicação sucinta dos algoritmos e estruturas de dados utilizados para resolver o problema. *(1 + 1 pontos)*
- Análise de complexidade dos principais algoritmos implementados. *(1 + 1 pontos)*
- Resultados de experimentos para avaliar empiricamente o desempenho dos algoritmos. Interprete os resultados *(3 + 3 pontos)*.