Metodologia Científica

Definição do projeto da Disciplina - 2015.1

Título: Investigação de diferentes técnicas de implementação de algoritmos de busca por textos.

Descrição

Você realizará uma investigação científica empírica visando a resolver o problema descrito a seguir.

Problema de negócio: Uma empresa oferece um serviço de busca por texto em grandes volumes de dados. A empresa deseja minimizar os recursos utilizados e, também, oferecer o melhor desempenho possível aos usuários. Porém, a empresa não sabe qual a técnica de implementação de algoritmo de busca por texto apresenta melhor desempenho.

Problema técnico: Analisar as diferentes técnicas de implementação de algoritmos de busca por textos em termos de desempenho. Para analisar o desempenho, podem ser usadas as seguintes métricas: tempo de execução, consumo de memória e número de operações realizadas[1].

Objetivo (template GQM)**: Analisar** técnicas de algoritmos de busca por textos **com a intenção de** comparação **com respeito ao seu** desempenho **do ponto de vista** de uma empresa que paga os recursos necessários e também dos seus usuários **no contexto** de grandes volumes de dados.

Milestone 1

Objetivo: Preparar artefatos necessários para realizar o experimento.

Atividades necessárias para realizar este *milestone*:

1. **Implementação de algoritmos:** Implementar os algoritmos de busca por textos usando o algoritmo de força bruta e mais duas entre as seguintes técnicas (total de 3 técnicas):
    - *Knuth-Morris-Pratt*;
    - *Boyer-Moore*;
    - *Boyer-Moore-Horspool*;
    - *Wagner-Fisher*;

[1] O número de operações realizadas refere-se à quantidade de vezes em que os laços foram executados.

- *Rabin–Karp*;
- *Aho–Corasick*.

Você também precisará implementar um programinha para ler arquivos de texto.
Dica: Existem diversos algoritmos disponíveis na *web* implementados com essas técnicas. Essas implementações podem ser usadas desde que não sejam protegidas por *copyright*. A linguagem de programação utilizada fica a seu critério.

2. **Preparar os dados para análise:** Você deve preparar arquivos de texto de tamanhos variados para: armazenar os textos que serão pesquisados e os textos onde serão realizadas as buscas. Cada arquivo deve conter apenas um texto a ser buscado ou um texto de busca. Prepare os arquivos de texto de busca contendo o maior volume de dados possível (sugestão: a partir de 500MB). Exemplo de um conjunto de dados**:** link.

3. **Selecionar uma forma de capturar as métricas:** Uma boa alternativa é o PIDSTAT[2] , mas existem vários.
Importante: Para as medições das métricas, considere apenas a parte de seu programa que se refere ao algoritmo de busca. Por exemplo, não inclua o tempo de leitura dos arquivos no tempo total de execução dos algoritmos de busca.
Dica: Você pode capturar o consumo de memória durante a execução de seu programa de tempos em tempos. Para calcular o consumo de memória final, você pode fazer uma média (ou mediana …) de "todos os consumos" coletados.

4. **Preparar *script* para coleta de dados:** Seu programa deve executar na linha de comando de uma máquina Linux. Você deve escrever um *script shell*[3] simples para realizar a coleta de dados. Chame o *script* de **`coletor_de_dados`**. Este *script* servirá não só para você coletar os dados mas também para uma simples verificação das saídas dos algoritmos (na fase de correção).
Importante: O seu *script* deve compilar (se for o caso) o código fonte disponibilizado.

   Os seguintes parâmetros são obrigatórios na entrada:
   - Técnica de implementação;
   - Caminho do arquivo com o texto que será buscado;
   - Caminho do arquivo com o texto onde a busca será realizada.

   A saída do coletor de dados deve informar os seguintes resultados na saída padrão: (i) resultado da busca (contém ou não contém o texto), (ii) tempo total de execução, (iii) consumo de memória e (iv) número de operações.

[2] PIDSTAT: http://linux.die.net/man/1/pidstat.
[3] Sobre Shell Script: http://pt.wikipedia.org/wiki/Shell_script.

Exemplo de entrada e saída:

Entrada: *./coletor_de_dados tecnica /home/exp/texto_buscado_1.txt /home/exp/teste/texto_busca_1.txt > out.txt*

Saída: *texto_buscado*:texto_buscado_1.txt *texto_busca*:nome_texto_busca_1.txt *resultado*: contem *tempo_execucao*: xx *consumo_memo*:yy *num_operacaoes*: zz

Deliverable: Um arquivo *Google Doc*, submetido via sistema de *peer assessment*, contendo:

1. Link para um arquivo armazenado no *Google Drive* denominado: *primeiroSobrenome.segundoSobrenome.tar.gz* contendo código fonte, *script shell* e os dados que serão utilizados no experimento.
2. As instruções de execução. Durante a correção, seu programa será executado com diferentes conjuntos de testes.

Critérios de avaliação para este *milestone*:

1. Implementação correta dos algoritmos.
2. Conjunto de dados para análise.
3. Testes.
4. Facilidade para rodar e testar.

Milestone 2

Objetivo: Planejamento do experimento.

Você deverá definir os elementos básicos de pesquisa: hipóteses, variáveis dependentes e independentes, fatores, níveis e *design* de experimento (lembre-se de randomização, pareamento ou blocagem (se houver) etc.). Lembre-se de descrever as suas hipóteses em termo das variáveis.

Sua pesquisa deve ter os seguintes fatores:

1. A técnica de implementação do algoritmo.
    a. Você deve usar ao menos 3 técnicas.
2. O tamanho do texto a ser buscado.
    a. Você pode considerar o número de caracteres do texto como unidade.
3. O tamanho do texto de busca.
    a. Você pode considerar o tamanho do arquivo como unidade.
4. Mais um fator interessante de sua escolha.

No planejamento, inclua:

- A obtenção de um modelo que explique os efeitos dos fatores, interações e erros experimentais.
- A determinação da alocação da variação;
- A descoberta do fator mais significativo;
- Uma reexecução do experimento, utilizando dois fatores apenas: a técnica e o fator mais significativo identificado no passo anterior (ou o segundo mais significativo, caso a técnica seja o fator mais significativo). Determinação da alocação da variação entre esses fatores e compare com o modelo original.
- Repetição da análise usando o tamanho do texto a ser buscado como fator-estorvo (fator desinteressante).

***Deliverable*:** documento no formato da investigação científica exemplo da disciplina (submetido via sistema de *peer assessment*).

Critérios de avaliação para este *milestone*:

1. Definição de hipótese(s) coerente(s) com a pergunta de pesquisa.
2. Escolha de bons fatores (os "botões") relativos ao contexto da pesquisa.
3. Escolha bem feita do *design* de experimento.
4. Uso adequado de randomização, blocagem e número de replicações.

Milestone 3

Objetivo: Execução, análise dos resultados e conclusões.

Execute seu experimento conforme o planejamento realizado no *milestone* 2. Realize testes estatísticos para verificar sua(s) hipótese(s) e analise os resultados finais obtidos.

***Deliverable*:** relatório com apresentação de resultados (lembre de usar gráficos adequados!) e conclusões. O relatório deve ser convincente, compreensível, conciso e bem-escrito. Deve ser submetido via sistema de *peer assessment*.

Critérios de avaliação para este *milestone*:

1. Identificação do modelo matemático experimental;
2. Validação do modelo;
3. Alocação da variação;
4. Significância estatística dos efeitos dos fatores;
5. Identificação do fator com maior efeito;
6. Qualidade das explicações para o comportamento visto nos gráficos do relatório;
7. Qualidade das explicações possíveis para a influência de cada fator nos resultados;
8. Apresentação de explicações plausíveis para os resultados finais obtidos no experimento, principalmente se a hipótese experimental não for aceita (hipótese nula

não rejeitada);
9. Muito importante: qualidade do português.

Dica: Comece a coletar os dados o quanto antes, pois a execução dos algoritmos pode levar bastante tempo.

Outros pontos importantes

Definição de equipes

Você pode formar pares ou trabalhar sozinho. Equipes não podem ter mais que 2 participantes então nem pergunte se pode.

Cronograma

	Data	**Peso na avaliação**
Milestone 1	4/6/2015	10%
Milestone 2	25/6/2015	40%
Milestone 3	9/7/2015	50%