UFMG/ICEx/DCC — Algoritmos e Estrutura de Dados 3

Trabalho Prático 5

Bacharelado — 1º Semestre de 2011

Esse trabalho prático tem por objetivo familiarizar o aluno com conceitos de localidade temporal/espacial e caracterização de dados.

1 Definição do problema

O ensino vem sofrendo mudanças decorrentes do avanço tecnológico. Uma delas é marcada pelo surgimento da educação à distância. Nessa modalidade de ensino o aluno muita das vezes assiste a aula pela Internet. Os vídeos-aulas podem ser assistidos pelos alunos várias vezes, além do fato de que eles podem movimentar o vídeo posicionando no momento desejado, seja retroagindo ou avançando.

Um vídeo-aula, neste caso, é transmitido por *chunks*. Cada *chunk* representa uma porção do vídeo que é identificada sequencialmente e possui um tamanho em Kbytes que não necessariamente é o mesmo para todos os chunks de um vídeo. Ou seja, um vídeo é composto de vários *chunks* e seu tamanho total é a soma dos tamanhos dos *chunks*.

No mundo perfeito a transmissão de um video ocorreria de forma sequencial lendo o primeiro *chunk* n do vídeo, em seguida o *chunk* $n+1$, $n+2$, ..., no entanto, na prática os alunos podem retroagir ou avançar o vídeo. Dado esse problema, como construir uma cache que melhore o desempenho da transmissão de vídeo em um sistema de educação à distância? Esse trabalho tem por objetivo a simulação de uma cache utilizando uma política de substituição de *chunks* que tenha um bom compromisso entre a quantidade de acertos (*chunks* encontrados na cache) e a ocupação do tamanho da memória disponibilizada para a cache. Para tal tarefa, esse trabalho está divido em dois momentos: Trabalho Prático 5 e Trabalho Prático 6.

2 Trabalho Prático 5

No Trabalho Prático 5 o que deve ser realizado é a primeira etapa da construção do simulador de cache em que o padrão de acesso aos vídeos e seus respectivos *chunks* são identificados através de uma caracterização. Para isso, o servidor de uma instituição de ensino com educação à distância foi monitorado ao longo de um determinado tempo. Essa monitoração resultou em um arquivo de log composto por três colunas: vídeo, *chunk* e tamanho do *chunk*. Cada linha desse arquivo de log representa a transmissão de um *chunk* de um determinado vídeo ao longo do tempo, esse arquivo de log está portanto ordenado pelo tempo, em que a linha x ocorreu primeiro que a linha seguinte $x+1$. Assume-se que existe um disco para armazenar cada vídeo no servidor e existe ainda uma controladora que acessa determinado disco de acordo o *chunk* requisitado. O tempo de operação da controladora é desprezível, porém o tempo de operação nos discos é proporcional a distância entre os *chunks*. Esses estão organizados em uma estrutura de dados sequencial de modo que a distância espacial entre dois *chunks* é dada pelo módulo da diferença de seus índices.

Considerando um exemplo de log(identificador do vídeo, índice do *chunk*, tamanho[1] do *chunk* em Kbytes):

1 0 10
2 0 5
1 1 8
1 0 10
2 0 5
1 5 13
1 0 10
2 4 7
2 1 50

[1] O tamanho do *chunk* será utilizado no Trabalho Prático 6.

Neste trabalho cada aluno terá um arquivo de log em que deve caracterizar as localidades de referência espacial e temporal e também informar a popularidade dos *chunks* dos vídeos produzindo as saídas especificadas na seção Formato de Saída. Deve-se, também, realizar uma análise dos resultados contendo os gráficos mostrados na seção Análise dos Resultados.

Localidade Temporal: É definida pela distância de pilha entre os *chunks* dos vídeos. Neste caso deve ser adicionado cada transação na pilha com vídeo e *chunk*. A cada novo acesso essa transação sobe para o topo da pilha e é calculado a distância da subida, exemplo:
Instante 1 - Pilha(do primeiro elemento para topo) = (1,0) - Distância da pilha = (sem distância)
Instante 2 - Pilha(do primeiro elemento para topo) = (1,0)(2,0) - Distância da pilha = (sem distância)
Instante 3 - Pilha(do primeiro elemento para topo) = (1,0)(2,0)(1,1) - Distância da pilha = (sem distância)
Instante 4 - Pilha(do primeiro elemento para topo) = (2,0)(1,1)(1,0) - Distância da pilha = 2
Instante 5 - Pilha(do primeiro elemento para topo) = (1,1)(1,0)(2,0) - Distância da pilha = 2,2
Instante 6 - Pilha(do primeiro elemento para topo) = (1,1)(1,0)(2,0)(1,5) - Distância da pilha = 2,2
Instante 7 - Pilha(do primeiro elemento para topo) = (1,1)(2,0)(1,5)(1,0) - Distância da pilha = 2,2,2
Instante 8 - Pilha(do primeiro elemento para topo) = (1,1)(2,0)(1,5)(1,0)(2,4) - Distância da pilha = 2,2,2
Instante 9 - Pilha(do primeiro elemento para topo) = (1,1)(2,0)(1,5)(1,0)(2,4)(2,1) - Distância da pilha = 2,2,2

Localidade Espacial: É definida pela módulo da diferença entre o índice do *chunk* atual e o índice do último *chunk* pertencente ao mesmo vídeo. No exemplo acima a localidade espacial por vídeo seria:
1 1,1,5,5
2 0,4,3
Distância espacial = 1,1,5,5,0,4,3

Popularidade: É definida pelo rank de *chunks* dos vídeos acessados ao longo de uma transmissão, ou seja, pela frequência de acesso.

No exemplo acima a popularidade seria:
(vídeo, chunk, frequência de acesso)
1 0 3
1 1 1
2 0 2
2 1 1
2 4 1

Os resultados dessa caracterização serão importantes na execução do Trabalho Prático 6, uma vez que elas podem ser utilizadas para definir as melhores estratégias de substituição da cache.

2.1 Análise dos Resultados

A seguir são mostrados exemplos de como os dados da localidade de referência temporal e espacial podem ser quantificados e analisados.

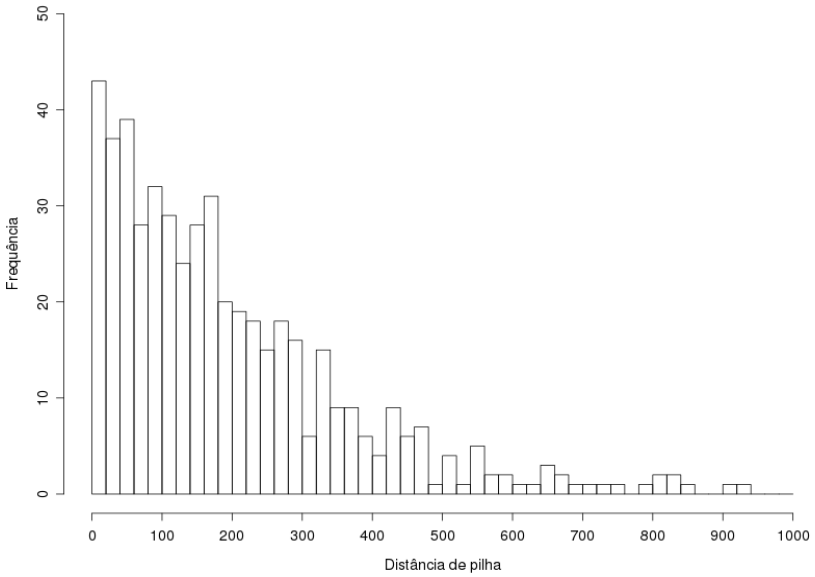
Os dados *A,B,C* e *D* estão armazenados sequencialmente (ordem lexcográfica) na memória do computador e em um determinado instante de tempo esses dados são acessados na seguinte ordem: **ABDBACCBDDAB**. As distâncias de pilha e as disposições de acesso desse padrão são demonstradas na tabela abaixo. Observe que a primeira e a última linha mostram as referidas distâncias, a segunda linha mostra o padrão de acesso e as linhas intermediárias simulam a movimentação da pilha, onde o dado que está sendo acessado está no topo da pilha.

Temporal	-	-	-	1	2	-	0	2	3	0	3	2
	A	B	D	B	A	C	C	B	D	D	A	B
	-	-	-	-	-	C	C	B	D	D	A	B
	-	-	D	B	A	A	A	C	B	B	D	A
	-	B	B	D	B	B	B	A	C	C	B	D
	A	A	A	A	D	D	D	D	A	A	C	C
Espacial	-	1	2	-2	-1	2	0	-1	2	0	-3	1

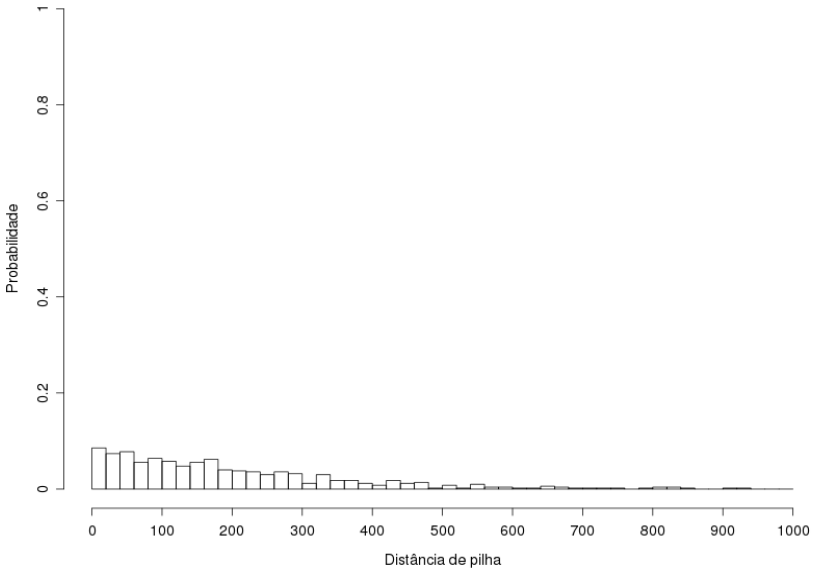
As distâncias acima poderiam ser sumarizadas com uma média, nesse caso a média da distância temporal seria 1,625 e espacial 0,090. Contudo deve-se levar em consideração a variação de cada distância, então um cálculo de desvio padrão ou intervalo de confiança da média pode indicar o quão representativa ela é. Logo tem-se um

desvio padrão 1,187 e 1,700 respectivamente para distância temporal e espacial, que indica a dispersão dos dados utilizados para calcular a média. Dividindo o desvio padrão pela média obtem-se o coeficiente de variação que indica a dispersão relativa dos dados, nesse exemplo, 0,730 e 18,7 para distância temporal e espacial, mostrando que essa última possui distâncias muita dispersas e a sua média é pouco representativa. Por outro lado, o intervalo de confiança fornece um erro para a média (somando ou subtraindo) com uma certa porcentagem de confiança, por exemplo, o erro das respectivas médias acima com 95% de confiança é +/-0,823 e +/-1,00.

Além de médias com desvio padrão, coeficiente de variação e erros, as distâncias de pilha ou espacial podem ser melhor visualizadas a partir de um histograma. Esse gráfico é construído dividindo o intervalo entre a menor e a maior distância em classes, em seguida é tabulado a frequência das distâncias para cada classe. Quando menor o tamanho das classes mais detalhado será o histograma. A seguir, tem-se um exemplo onde acesso aos dados *A,B,C* e *D* foram registrados em um instante de tempo maior, obtendo-se uma maior quantidade de distâncias de pilha. As frequências dessas distâncias são apresentadas pelo histograma abaixo e observe que as distâncias com valores até 20 (classe 0 a 20) foram as que apresentaram maior frequência (cerca de 43).



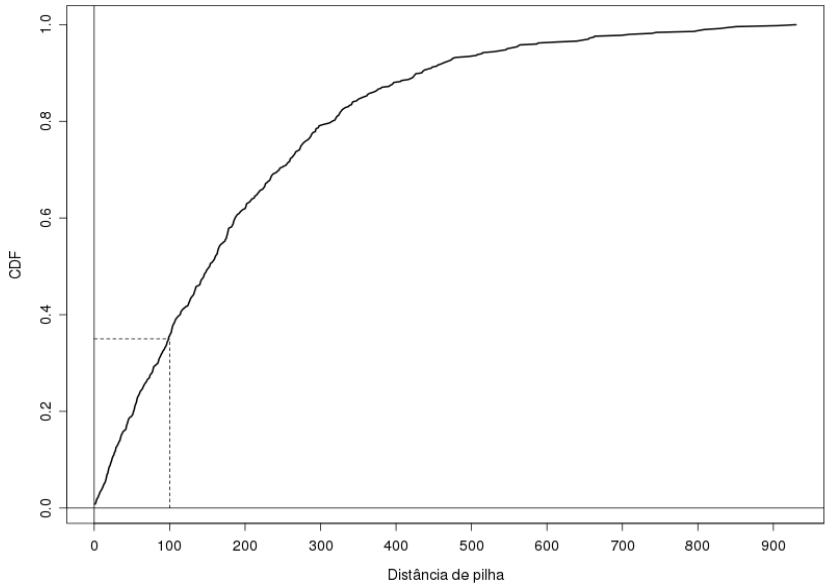
O eixo y do histograma, ao invés de representar a frequência em que as distâncias ocorrem, pode representar o total das frequências na escala de 0 a 1 e as barras verticais podem representar a proporção da frequência nessa escala. Desse modo o histograma exibe a função de distribuição de probabilidade de massa (PMF) das distâncias de pilha e os pontos altos no eixo y indicam as distâncias que ocorrem com maior probabilidade.



As probabilidades da PMF ainda podem ser acumuladas e formar um gráfico da função de distribuição de probabilidade cumulativa (CDF). Em teoria da probabilidade, uma CDF descreve completamente a distribuição da probabilidade de uma variável aleatória de valor real X. Para cada valor de x, a CDF é dada por:

$$F(x) = P(X \leq x)$$

Assim, para um dado valor x uma CDF indica a probabilidade de quaisquer outros valores X serem menores ou iguais a x. No gráfico a seguir as distâncias de pilha analisadas anteriormente são expressas por uma CDF. Na leitura desse gráfico observa-se que cerca de 35% (0.35 no eixo y) dos dados observados tem distância de pilha menor ou igual a 100 (valor no eixo x correspondente a 0.35 no eixo y).



Os exemplos e gráficos acima mostraram resumidamente que há mais de uma forma para visualizar e representar um conjunto de dados (as distâncias). Para aprender mais sobre esses conceitos o aluno pode consultar as referências abaixo ou ainda consultar a documentação de ferramentas como MS Excel, BrOffice, MatLab, R e GNUPlot que podem inclusive ser utilizadas para sumarização de dados ou construção de gráficos neste trabalho. Caso seja necessário, mais informações sobre esses gráficos serão disponibilizadas via moodle. Neste trabalho prático (TP5), o aluno deverá apresentar na documentação a caracterização das localidades de referência temporal e espacial do log de vídeo utilizando os gráficos mostrados nessa seção e relatar as conclusões que podem ser obtidas a partir desses gráficos.

Referências:

- Introdução a estatística 10. ed. Mario F. Triola
- Probabilidade: um curso moderno com aplicações - 8.ed. Sheldon M. Ross.
- Introduction to probability and statistics for engineers and scientists. Sheldon M. Ross.

2.2 Formato de Entrada

O arquivo de log possui várias linhas em que cada linha apresenta uma transação realizada e é composta por três colunas:

- Coluna 0: Identificador do vídeo;
- Coluna 1: Índice do *Chunk* de vídeo;
- Coluna 2: Tamanho do *chunk* em Kbytes;

2.3 Formato de saída

Devem ser gerados três arquivos textos com os seguintes nomes: localidadetemporal.txt, localidadeespacial.txt e popularidade.txt. Cada arquivo deve seguir o seguinte formato:

localidadetemporal.txt: Cada linha do arquivo deve representar a localidade de referência temporal no instante em que um *chunk* é acessado, com as distâncias de pilha separadas por vírgulas. Nos primeiros instantes em que não houve distâncias entre acessos, as linhas devem conter um "()". Não acrescente outros símbolos ou espaços. Exemplo:

()
()
()
2
2,2
2,2
2,2,2
2,2,2
2,2,2

localidadeespacial.txt: Cada linha do arquivo deve representar a localidade espacial entre os *chunks* de um vídeo, em que o primeiro item da linha é definido pelo índice do vídeo seguido por **um** espaço e as distâncias espaciais dos *chunks* lidos separadas por vírgulas. Na última linha do arquivo as distâncias espaciais de todos os vídeos devem ser mescladas. Não acrescente outros símbolos ou espaços. Exemplo:

1 1,1,5,5
2 0,4,3
1,1,5,5,0,4,3

popularidade.txt: Cada linha do arquivo deve representar a popularidade dos vídeos, em que a primeira coluna é definida pelo identificador do vídeo, a segunda coluna pelo índice do *chunk* e a terceira coluna a popularidade. As colunas devem ser separadas por vírgulas. Não acrescente outros símbolos ou mais de um espaço. Exemplo:

1,0,3
1,1,1
2,0,2

3 O que deve ser entregue:

Documentação:

- Introduza o problema apresentado.
- Explique os principais algoritmos utilizados para resolver o problema e a sua complexidade de tempo e espaço.
- Explique brevemente as decisões e módulos de sua implementação.
- Analise os dados que foram impressos pelo programa e relate quais conclusões podem ser obtidas a partir deles.
- Analise os gráficos construídos com informações sobre localidade de referência temporal e espacial e relate quais conclusões podem ser obtidas a partir deles.
- Faça uma conclusão breve sobre os resultados obtidos acerca de todos os dados e gráficos gerados nesse TP.
- Use o *template* de documentação presente no minha.ufmg.
- A documentação não pode exceder 10 páginas.

Código:

- O código fonte do trabalho deve ser submetido para compilação e execução em ambiente Linux, tendo como padrão os computadores dos laboratórios de graduação do DCC;
- Deve ser escrito na linguagem C (trabalhos implementados em outras linguagens como C++/Java/Python e outras não serão aceitos);
- As estruturas de dados devem ser alocadas dinamicamente e o código ser modularizado (ou seja, dividido em múltiplos arquivos fonte e fazendo uso de arquivos cabeçalho - .h);
- O utilitário Make deve ser utilizado para compilar o programa;
- Os arquivos gerados pelo programa devem seguir estritamente o formato da especificação, caso contrário o resultado será considerado errado;
- O arquivo executável deve ser chamado de **tp5** e deve receber como parâmetro apenas o nome do arquivo de entrada de dados. **Não serão aceitos outros nomes de executáveis além dos mencionados**.
- Faça seu código de forma legível;

Entrega:

- Data de entrega : 22/06/2011 .
- Submissão: a documentação, o arquivo pdf de gráficos e o código do trabalho devem ser submetidos ao minha.ufmg. Para isso compacte os dois (formato tar.gz) e faça a submissão. Teste seu arquivo compactado antes de enviá-lo.
- Apenas a documentação também deve ser entregue impressa na secretaria do DCC. Não coloque nos escaninhos dos professores. A documentação deve ser entregue para a secretária e então colocada no envelope de AEDS3.
- Será postada uma planilha no Moodle sobre a entrevista do trabalho, leia-a e siga as orientações para o agendamento da sua entrevista.

Distribuição dos pontos:

- Execução: 50%
- Documentação: 50%

Será adotado média harmônica entre a pontuação obtida na execução e na documentação do TP, o que implica em valor zero caso alguma das partes não seja apresentada.