Universidade Estadual de Maringá
Centro de Tecnologia
Departamento de Informática

Prof: Nilton Luiz Queiroz Junior
Disciplina: Programação Concorrente (5205 – Turma 31)

Segundo Trabalho Prático

Objetivo: O trabalho tem como objetivo a implementação de uma aplicação paralela, utilizando troca de mensagens e a elaboração de um artigo, fazendo análises de desempenho da aplicação implementada.

Instruções:

- O trabalho poderá ser feito em dupla ou individual;
- A implementação da aplicação deverá ser feita na linguagem C/C++, Java ou Python utilizando a biblioteca MPI;
  - Caso a dupla deseje utilizar outra linguagem, a qual exista uma implementação da MPI, deverá conversar com o professor;
- O trabalho tem valor de 0 a 10 e corresponde à segunda avaliação periódica da disciplina;
- Deverão ser entregues os seguintes itens:
  1. O código fonte do trabalho;
     - Tanto da versão sequencial quanto da paralela;
       - A versão sequencial pode ser a mesma do primeiro trabalho;
  2. Um passo a passo de como compilar e executar o programa;
     - Como deve ser fornecida a entrada para o programa, qual sistema operacional a ser usado, e caso seja usado algum outro recurso além da biblioteca MPI, informar qual recurso foi usado, como instalar e como configurar, etc;
  3. O artigo, no modelo para a publicação de artigos da SBC, de 4 a 8 páginas, fazendo analise de desempenho composto por:
     - Titulo;
     - Resumo
       - Não é obrigatório o *Abstract* (resumo em inglês);
     - Introdução;
     - Referencial teórico;
       - Descrever superficialmente técnicas que foram usadas na paralelização do algoritmo:
         - Por exemplo:
           - Descrever os conceitos de passagem de mensagem e os demais mecanismos das operações utilizadas;
     - Desenvolvimento;
       - Explicar como funciona a aplicação e como ela foi paralelizada;
         - Como foi feita a divisão do trabalho;
         - Caso tenha sido usado algum paradigma de interação entre processos, comentar sobre o paradigma e como ele foi aplicado na implementação do algoritmo;
       - Entre outros detalhes;
     - Ambiente experimental e experimentos realizados;
       - Descrever hardware e sistema operacional utilizado;
         - Caso a equipe configure uma rede, deve informar as especificações de cada nó da rede (Sistema Operacional, Hardware, etc);

- Descrever detalhes dos experimentos que foram realizados;
  - Tamanho das entradas;
  - Quantidade de *processos;*
    - Caso seja feito em rede, informar quantos processos foram utilizados por nó da rede;
  - Parâmetros do algoritmo;
  - Quantas execuções foram realizadas para cada variação do experimento;
- Análise e discussão dos resultados;
  - Extrair métricas tais como *speedup*, entre outras vistas em sala, e discutir os resultados com base nas métricas obtidas tais métricas;
    - Colocar gráficos (ou tabelas) com *speedup*, e outras métricas consideradas importantes pela equipe, por quantidade de *processos*;
    - Tentar encontrar relações entre os resultados obtidos e a teoria vista em sala;
    - Em casos de resultados que não tem o comportamento esperado, tentar levantar hipóteses para justificar;
      - Não é necessário realizar experimentos para justificar tais hipóteses, apenas aponta-las como possível causa dos problemas, e deixar tal investigação em aberto como trabalhos futuros.
- Conclusões;

- Link para o download dos templates de modelo de artigo da SBC:
  - http://www.sbc.org.br/documentos-da-sbc/category/169-templates-para-artigos-e-capitulos-de-livros
- Os itens acima deverão ser entregues via moodle, em um arquivo compactado que deve seguir o padrão:
  - NomeAluno1_RAXXXXX_NomeAluno2_RAXXXXX;
    - O formato do arquivo para submissão poderá ser .ZIP, .RAR ou .TAR
- A data limite para a entrega do trabalho será combinada em sala de aula;

Descrição: Neste trabalho deverá ser implementado o algoritmo K-means, fazendo a paralelização dos cálculos realizados sobre os exemplos da base de dados. Serão disponibilizadas 5 bases
A localização dos arquivos de base fica a critério da equipe, ou seja, pode-se assumir que os arquivos de base estão presentes em todos os nós da rede, ou então que um único processo irá ler e distribuir o arquivo pela rede.
Cada base tem uma quantidade diferente de atributos, porém todas elas seguem o mesmo padrão:

- Cada linha representa um exemplo;
- Todos os atributos de cada exemplo são números inteiros;
- Os atributos são todos separados por vírgula;

Além das bases, também serão disponibilizados os arquivos contendo os atributos dos centroides iniciais, que seguem o mesmo padrão.
As bases tem a seguinte nomenclatura:

int_base_<numero_de_atributos>.data

Os arquivos com os centroides tem a seguinte nomenclatura:

int_centroides_<numero_de_atributos>_<numero_de_centroides>.data

Por exemplo:

int_base_256.data

Esses arquivo indica que a base tem 256 atributos, ou seja, são 256 números reais os quais deverão ser usados para a clusterização.

int_centroides_256_20.data

Esse arquivo indica que existem 20 centroides com 256 atributos (consequentemente, estes centroides devem ser utilizados junto a base de 256 atributos).

Para um melhor entendimento do algoritmo K-means, a seguir serão apresentados alguns conceitos de aprendizagem de máquina e uma pequena descrição do algoritmo.

Um dos principais objetivos da maioria dos algoritmos de aprendizagem de máquina é classificar exemplos baseados em seus atributos. Dentro da área de aprendizagem de máquina existem algoritmos que tentam agrupar os exemplos levando em consideração treinamento, os quais dependem de exemplos previamente classificados e são conhecidos como algoritmos de classificação supervisionada. Além destes, existem outros algoritmos que tentam agrupar os exemplos, sem a necessidade de exemplos prévios, os quais são chamados de algoritmos de classificação não supervisionada.

Ambas os tipos de algoritmos levam em consideração um conjunto de características extraídos dos exemplos, tais características são chamadas de atributos.

Métodos de aprendizagem supervisionados necessitam de um atributo denominado classe, o qual determina o que se irá aprender. Estes métodos usam a classe de exemplos anteriores para predizer a classe de um novo exemplo. Já métodos não supervisionados tentam agrupar os exemplos que tem maior nível de similaridade, pois grupos similares podem ser vistos como potenciais classes.

O algoritmo K-means é um algoritmo de classificação não supervisionada, e a ideia por trás de tal algoritmo é processar um conjunto de exemplos, de maneira a agrupá-los em K grupos distintos.

Porém, o algoritmo K-means não é capaz de decidir a quantidade de agrupamentos a serem gerados, então, a quantidade de agrupamentos é um parâmetro do algoritmo.

A ideia geral por trás do algoritmo K-means é utilizar K centroides, os quais serão as referências para cada um dos agrupamentos (os centroides **podem não** ser exemplos que estão nos agrupamentos).

Desta forma, o algoritmo irá executar e agrupar os elementos, associando cala exemplo ao centroide mais próximos, dando a todos eles o mesmo rotulo. Após rotular os exemplos é necessário reposicionar cada um dos centroides. O algoritmo irá então se repetir até que nenhum centroide sofra qualquer alteração. Um pseudocódigo para o algoritmo K-means é mostrado na Figura 1, sendo B a base de dados e C o conjunto de centroides (contendo k centroides).

O calculo de distância e o procedimento de recalcular a posição dos centroides podem ser feitos de diversas maneiras. A mais comum para o calculo de distâncias é a distância euclidiana, exibida a seguir:

Distancia Euclidiana entre dois pontos:

$$D(P,Q)=\sqrt{\sum_{i=1}^{n}(P_i-Q_i)^2}$$

Universidade Estadual de Maringá
Centro de Tecnologia
Departamento de Informática

Prof: Nilton Luiz Queiroz Junior
Disciplina: Programação Concorrente (5205 – Turma 31)

```
K-means(B, C)
        Repita
                Para cada exemplo E ∈ B
                        E.rotulo = Rótulo do centroide Ci ∈ C mais próximo de E
                Para cada Centroide Ci ∈ C
                        Recalcule as coordenadas de Ci
        Até que não exista alteração em C
```

Figura 1: Pseudo-código para o algoritmo K-Means

Já o reposicionamento do centroide, é mais comumente calculado por meio das médias dos exemplos que possuem o mesmo rotulo do centroide. Dessa forma, para todos os pontos que possuem o mesmo rótulo do centroide, obtém-se a média de cada um dos atributos, e cada uma destas médias é o novo "atributo" do centroide.
Por exemplo:

Suponha uma base onde cada exemplo possui 4 atributos. Suponha também que temos um dos centroides C_x com as seguintes coordenadas:(20,37,43,58). Além disso, apenas três pontos tem C_x como o ponto mais próximo, e são eles:

$E_1 = (20,38,43,60)$; $E_2 = (18,33,45,62)$; e $E_3 = (25,37,44,64)$

Com isso, o novo centroide será calculado da seguinte maneira:

O primeiro "atributo" do centroide será realizado pela média aritmética do primeiro atributo dos pontos E_1, E_2 e E_3: (20 + 18 + 25)/3 = 21
O segundo "atributo" do centroide será: (38 + 33 + 37)/3 = 36
O terceiro "atributo" do centroide será dado por: (43 + 45 + 44)/3 = 42
E o quarto "atributo" será dado por: (60 + 62 + 64)/3 = 62

Portanto, o centroide C_x passará então a ser: (21, 36, 42, 62).

Obs: Devido à sensibilidade do algoritmo, quando um exemplo possuir a mesma distância para dois ou mais centroides, tanto na versão paralela quanto na sequencial, deve ser determinado o mesmo critério de desempate, para que não exista o risco de quantidades de iterações nas versões sequencial e paralela.

Problemas com Trabalhos <u>COPIADOS</u>:

Quem copiar terá o trabalho anulado (zerado), seja de outra dupla ou da internet.
Quem fornecer a cópia também terá o trabalho anulado (zerado).

Referências para o algoritmo de agrupamento:

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
Capittulo 8.2.

Ian H. Witten and Eibe Frank. 2005. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
Capitulo 4.8.