UNIVERSIDAD EAFIT

ST0263: Tópicos Especiales en Telemática, 2014-2

Trabajo Final – Opción 2

Procesamiento Paralelo - distribuido: Sistemas de recomendación basado en la Correlación de Pearson,

Contexto:

Un Sistema de Recomendación (SR), es una funcionalidad de un sistema de información que permite realizar recomendaciones de productos, contenidos, objetos, o muchas otras cosas a los usuarios del sistema (dependerá del tipo de aplicación o sistema). El pionero en estos sistemas fue Amazon. Recuerde que cuando ud realiza consultas o navega por el sitio de Amazon, el sistema le recomienda o sugiere productos basados en sus compras anteriores, gustos similares con otros usuarios, preferencias, etc.

Un servicio particular que utiliza los sistemas de recomendación son los de video on demand, es decir, un usuario solicita peliculas bajo demanda (ver Netflix o la Tienda de video de UNE).

Existen muchas técnicas o métodos para implementar sistemas de recomendación en Video, el que utilizarémos aca, se basa en el comportamiento histórico de las peliculas que han visto los usuarios y la calificación (rating) que han hecho los usuarios de las peliculas.

La estructura y secuencia general de este tipo de recomendador funciona de la siguiente forma:

IN: archivo de logs de usuarios (usuarios-peliculas.txt)
1. Cargar los logs de usuarios en la matriz "MatrixUI[Mx,My]" con las dimensiones Mx (número de peliculas del servidor), My (número de usuarios en el servidor). Cada entrada de la matriz$_{i,j}$ es el rating o calificación que realizó el usuario i sobre la pelicula j.

2. Calcular la matriz de correlación entre usuarios (MatrixCorr[My,My]). Es una matriz de My por My. Cada entrada de la matriz MatrixCorr$_{k,l}$ mide el nivel de similitud en preferencias entre el usuario k y el usuariol l. Cada valor de MatrixCorr$_{k,l}$ es un valor entre [-1,1]. Donde 1 indica una coincidencia total entre preferencias. -1 gustos totalmente diferentes. 0 no se puede decir nada.

3. Calcular la matriz de recomendación (SR[m,My]). Se realizará el ordenamiento por filas de la matriz MatrixCorr (exceptuando la diagonal) y se

seleccionarán los m valores más cercanos a 1. Normalmente se puede definir m entre 5 y 10.

Los pasos 1,2 y 3 se consideran preparatorios para ejecutar la función de recomendación. Los siguientes pasos los realiza cuando va a realizar la recomendación.

Vamos a suponer que se logea el usuario n en el sistema (n entre 1 y My).

4. Ingresa al sistema el usuario n.

5. Para recomendar p peliculas al usurio n, se buscan los usuarios similares a n de acuerdo a la matriz SR[h,n], en orden descendente de correlación se buscan las peliculas que no ha visto n con mayor rating del usuario h (h: 1..m). Ver el siguiente extracto de un artículo de un sistema de recomendación:

(1) Find users who are similar to user a (the neighbours of a).
(2) Select the "nearest" neighbours of a, i.e. select the most similar set of users to user a.
(3) Recommend items that the nearest neighbours of a have rated highly and that have not been rated by a.

A continuación se presenta un extracto de un artículo donde aparece el calculo de la correlación de pearson:

Standard statistical measures are often used to calculate the similarity between users in step 1 (e.g. Spearman correlation, Pearson correlation, etc.).[32] In this work, similar users are found using the Pearson correlation coefficient formula (7.1):

$$\text{corr}_{a,u} = \frac{\sum_{i=1}^{m} (r_{a,i} - \bar{r_a}) \times (r_{u,i} - \bar{r}_u)}{\sqrt{\sum_{i=1}^{m} (r_{a,i} - \bar{r_a})^2} \times \sqrt{\sum_{i=1}^{m} (r_{u,i} - \bar{r_u})^2}} \qquad (7.1)$$

where $\text{corr}_{a,u}$ is the correlation value between users a and u (a value in the range $[-1, 1]$) for m items rated by users a and u, $r_{a,i}$ is the rating given by user a to item i, $r_{u,i}$ is the rating given by user u to item i, $\bar{r}_a$ and $\bar{r}_u$ are the average ratings given by users a and u, respectively.

El mayor tiempo computacional de este algoritmo se presenta en el calculo de la matriz de correlación (MatrixCorr) y el ordenamiento de la matriz MatrixCorr por filas para seleccionar los m valor mayores.

En un escenario real de uso de este sistema, podríamos estar hablando de dimensiones de la matriz de usuarios-peliculas del orden de Mx=10.000 peliculas. My=100.000 usuarios. Como podrá ver, esta matriz supone ser muy dispersa, sin embargo para este ejercicio no se considerará esta dispersión.

Requerimiento:

Realizar el diseño e implementación de un algoritmo paralelo que puede ser ejecutado en un cluster MPI, que permita disminuir el tiempo de procesamiento para la generación de la matriz SR (debe ejecutarse la correlación de pearson y el ordenamiento por filas, finalmente para seleccionar los m valores mayores).

Se tomará como entrada una matriz de usuarios-peliculas, de dimensiones paramétricas Mx,My. Los valores de ratings están entre 0 y 5. 0 no vista y 1-5 la opinión de la pelicula. 1 muy mala. 5 muy buena.

La salida será la matriz SR[m,My].

Para simular los valores de ratings, dados Mx,My,m. Llenar aleatoriamente la matriz: matrizUI[Mx,My].

Comprobar si el algoritmo serial (sr-serial.c) es más lento que el algoritmo paralelo (sr-paralelo.c) y cual sería la eficiencia del algoritmo paralelo.

Síntesis de actividades a realizar:

1. Entender mejor el algoritmo de correlación de pearson para los sistemas de recomendación.
2. Realizar el diseño del algortimo paralelo.
    a. Definir diferentes (al menos una) estrategias de división de este algoritmo para llevarlo a paralelo. Analizar los casos de particionamiento por Datos o por Funcionalidad. Utilizar la estrategia PCAM. Explicitamente definir el grafo de dependencias, el grado de concurrencia y la estrategia de comunicación de datos (punto a punto: sync/async, buffered/non-buffered. Colectivas: broadcast, reduce, scatter, etc).
3. Realizar la estimación de rendimiento del algortimo paralelo de forma analitica.
4. Codificar el algoritmo en C en forma secuencial, tomar el tiempo de procesamiento.
5. Realizar diferentes experimentos que permita comprobar el tiempo de ejecución en producción del programa paralelo y documentar el análisis

Instalar, Configurar y Utilizar un cluster MPI ha ser instalado en el Data Center Academico.

Se proveeran las n maquinas virtuales para la realización del proyecto