Escuela Profesional de Ciencias de la Computación

Un motor de recomendación de filtrado colaborativo utilizando Openmp

Presentado a:
Dra. Anamaria Cuadros.

Presentado por:
Lipa Urbina, Edson V.
Condori, Christian.

Índice

1. Introducción

En el mundo del big data, el sistema de recomendación se está haciendo cada vez más popular. La razón es que esta herramienta automatizada conecta al comprador con los productos más adecuados para comprar al correlacionar los contenidos del producto y los comentarios expresados. Una de las técnicas prevalecientes más destacadas del motor de recomendación es el filtrado colaborativo. Depende solo de las acciones pasadas del usuario, como la transacción pasada o la retroalimentación del artículo. Los algoritmos tradicionales de filtrado colaborativo, como el enfoque de vecindad y los modelos de factores latentes, suelen tener tres problemas principales. En primer lugar, el Problema, que está básicamente relacionado con el desglose de los recomendadores que no pueden inferir preferencias, especialmente para los nuevos usuarios para los cuales no tiene información suficiente. En segundo lugar, *Scalability* que se puede definir como la capacidad del Recomendador de producir recomendaciones en tiempo real o casi en tiempo real para conjuntos de datos a gran escala. Por último, *Sparsity* de la matriz de clasificación de elementos de usuario, ya que la mayoría de los usuarios activos solo puntuarán pocos elementos de todos los elementos totales. Para resolver el problema de las recomendaciones, muchos investigadores probaron diferentes enfoques como la agrupación en clústeres y la creación de recomendaciones basadas en características utilizando *tag*.

2. Objetivo

El objetivo de esta invetigación es el de desarrollar un sistema de recomendación que pueda ser eficiente al momento de retornar una consulta teniendo en cuenta también el espacio que se utiliza para el procesamiento de los datos, ya que al tener una gran cantidad de datos para procesar también se utiliza memoria para analizar los datos

3. Conceptos Generales

3.1. Metricas de Distancia

3.1.1. Distancia Euclidiana

La distancia Euclidiana o métrica Euclidiana es la distancia entre dos puntos que uno mediría con una regla. Es la longitud del segmento de línea que los conecta.**?**

En un plano donde $p1 = (x1, y1)$ y $p2 = (x2, y2)$ esta es:

$$dist(p1, p2) = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

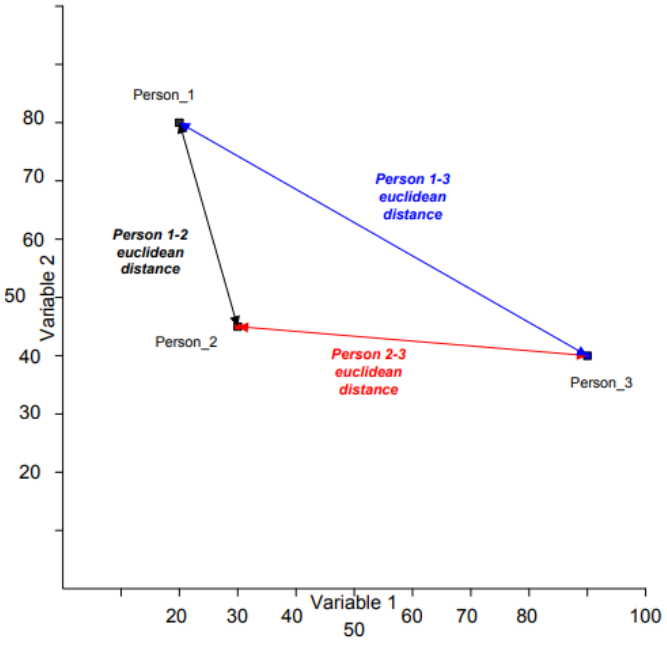

Figura 1: Distancias entre tres puntos

En N dimensiones la distancia Euclidiana entre dos puntos p y q seria:

$$dist(p, q) = \sqrt{\sum_{i=1}^{N}(pi - qi)^2}$$

La distancia euclidiana es el uso más común de la distancia. En la mayoría de los casos, cuando las personas dicen acerca de la distancia, se referirán a la distancia euclidiana. La distancia euclidiana también se conoce como simple distancia. Cuando los datos son densos o continuos, esta es la mejor medida de proximidad.

3.1.2. Distancia de Manhattan

La distancia entre dos puntos medidos a lo largo de los ejes en ángulos rectos.**?** En un plano donde $p1 = (x1, y1)$ y $p2 = (x2, y2)$ esta es:

$$dist(p1, p2) =\mid x1 - x2 \mid + \mid y1 - y2 \mid$$

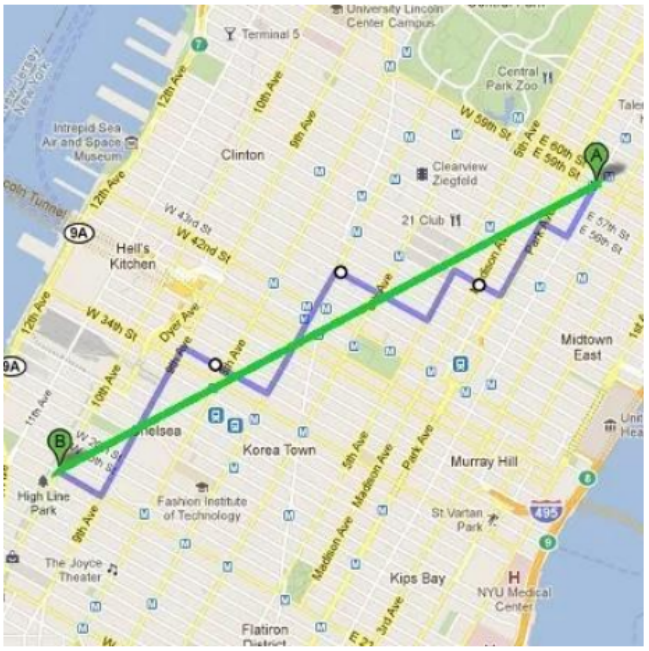

Figura 2: Distancias Manhattan(morado) y Euclidiana(verde)

Esta métrica se ve menos afectada por valores atípicos que las metrías euclidianas.

Esto es fácilmente generalizado a dimensiones más altas. La distancia de Manhattan a menudo se usa en circuitos integrados donde los cables solo corren paralelos al eje X o Y.Esta métrica de distancia de Manhattan también se conoce como longitud de Manhattan, distancia rectilínea, distancia L1 o norma L1, distancia de bloque de ciudad, distancia L1 de Minkowski, métrica de taxi o distancia de bloque de ciudad. La distancia de manhattan se basa en la distancia de valor absoluto, a diferencia de la distancia de error cuadrado. En la práctica, debes obtener resultados similares la mayor parte del tiempo. La distancia de valor absoluto debería dar resultados más robustos, mientras que Euclidean se vería influenciada por valores inusuales.

3.1.3. Similitud de Coseno

Es la medida de similitud existente entre dos vectores en un espacio que posee un producto interior con el que se evalúa el valor del coseno del ángulo comprendido entre ellos.**?**

$$Cos(A,B) = \frac{\sum_{i=1}^{N}(AiBi)}{\sqrt{\sum_{i=1}^{N}(Ai)^2}\sqrt{\sum_{i=1}^{N}(Bi)^2}}$$

Esta función trigonométrica proporciona un valor igual a 1 si el ángulo comprendido es cero, es decir si ambos vectores apuntan a un mismo lugar. Cualquier ángulo existente entre los vectores, el coseno arrojaría un valor inferior a uno.

Si los vectores fuesen ortogonales el coseno se anularía, y si apuntasen en sentido contrario su valor sería -1. De esta forma, el valor de esta métrica se encuentra entre -1 y 1, es decir en el intervalo cerrado [-1,1].

Esta distancia se emplea frecuentemente en la búsqueda y recuperación de información representando las palabras (o documento) en un espacio vectorial. En minería de textos se aplica la similitud coseno con el objeto de establecer una métrica de semejanza entre textos. En minería de datos se suele emplear como un indicador de cohesión de clusteres de textos. La similitud coseno no debe ser considerada como una métrica debido a que no cumple la desigualdad triangular.

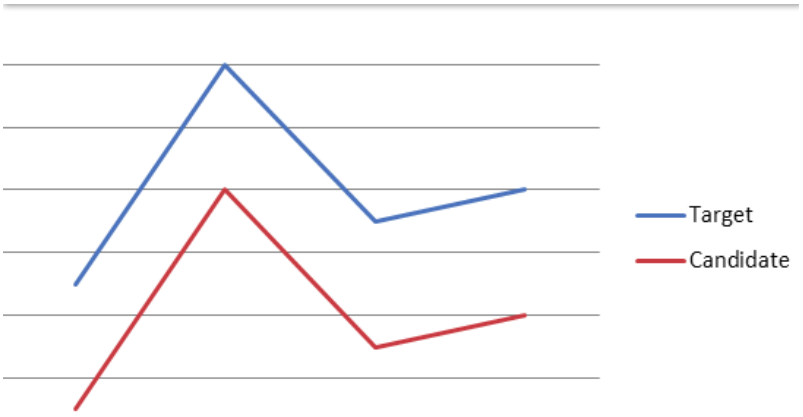

Figura 3: Perfil similar

Este método de similitud se utiliza para encontrar lugares que tienen las mismas características pero quizá una mayor o menor escala.
Una ventaja de la similitud de coseno es su baja complejidad , especialmente para vectores dispersos : solo deben considerarse las dimensiones que no son cero.

Otros nombres de similitud de coseno son la similitud de Orchini y el coeficiente de congruencia de Tucker ; La similitud de Ochiai es la similitud de coseno aplicada a datos binarios.

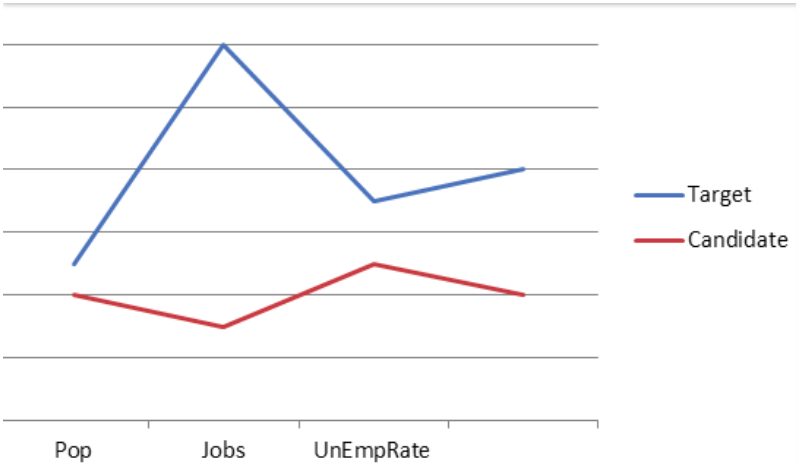

Figura 4: Perfil diferente

Medida del coseno suave Un coseno suave o (similitud "blanda") entre dos vectores considera similitudes entre pares de características. La similitud de coseno tradicional considera que las características del modelo de espacio vectorial (VSM) son independientes o completamente diferentes, mientras que la medida de coseno suave propone considerar la similitud de las características en VSM.
Por ejemplo, en el campo del procesamiento de lenguaje natural (PNL), la similitud entre las características es bastante intuitiva. Las características tales como palabras, n -grams o n -grams sintácticos pueden ser bastante similares, aunque formalmente se consideran características diferentes en el VSM. Por ejemplo, las palabras "jugar" y "game"son palabras diferentes y, por lo tanto, se asignan a diferentes puntos en VSM; sin embargo, están semánticamente relacionados.

$$SoftCos(A,B) = \frac{\sum_{i,i}^{N}(SijAiBi)}{\sqrt{\sum_{i,j}^{N} SijAiAj}\sqrt{\sum_{i=1}^{N} SijBiBj}}$$

- Donde **Sij** = Similitud(Caracteristica i, Caracteristica j)

3.1.4. Coeficiente de Correlación de Pearson

El coeficiente de correlación producto-momento de Pearson (o coeficiente de correlación de Pearson) es una medida de la fuerza de una asociación lineal entre dos variables y se denota con r. Básicamente, una correlación de Pearson intenta trazar una línea de mejor ajuste a través de los datos de dos variables, y el coeficiente de correlación de Pearson, r, indica qué tan lejos están todos estos puntos de datos a

esta línea de mejor ajuste (es decir, que tan bien, los de datos se ajustan a este nuevo modelo / línea de mejor ajuste).**?**

El coeficiente de correlación de Pearson, r, puede tomar un rango de valores de +1 a -1. Un valor de 0 indica que no hay asociación entre las dos variables. Un valor mayor que 0 indica una asociación positiva; es decir, a medida que aumenta el valor de una variable, también lo hace el valor de la otra variable. Un valor menor que 0 indica una asociación negativa; es decir, a medida que aumenta el valor de una variable, el valor de la otra variable disminuye. Esto se muestra en la siguiente figura 5

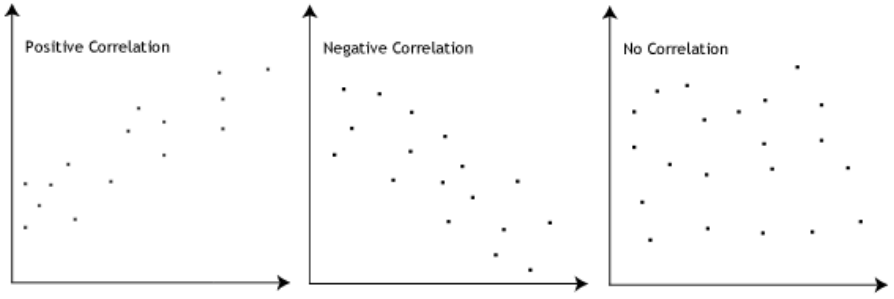

Figura 5: Tipo de correlaciones de pearson

Cuanto más fuerte sea la asociación de las dos variables, más cercano será el coeficiente de correlación de Pearson, r, a +1 o -1, dependiendo de si la relación es positiva o negativa, respectivamente. Alcanzar un valor de +1 o -1 significa que todos sus puntos de datos están incluidos en la línea de mejor ajuste; no hay puntos de datos que muestren ninguna variación fuera de esta línea. Los valores de r entre +1 y -1 (por ejemplo, r = 0.8 o -0.4) indican que existe una variación alrededor de la línea de mejor ajuste. Cuanto más cercano sea el valor de r a 0, mayor será la variación alrededor de la línea de mejor ajuste.

La correlación producto-momento de Pearson no tiene en cuenta si una variable se ha clasificado como una variable dependiente o independiente. Se trata a todas las variables por igual.Además Las variables se pueden medir en unidades completamente diferentes. Por ejemplo, se podría correlacionar la edad de una persona con sus niveles de azúcar en la sangre.

Es importante darse cuenta de que el coeficiente de correlación de Pearson, r, no representa la pendiente de la línea de mejor ajuste. Por lo tanto, si obtiene un coeficiente de correlación de Pearson de +1, esto no significa que por cada aumento

de unidad en una variable haya un aumento de unidad en otra. Simplemente significa que no hay variación entre los puntos de datos y la línea de mejor ajuste.**?**

3.2. Algoritmos

3.2.1. K-NN

El método de los k vecinos más a cercanos (k-nearest neighbors, K-NN) es un método de clasificación supervisada que sirve para estimar la función de densidad de las predictoras por cada clase.

Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento **x** pertenezca a la clase a partir de la información proporcionada por el conjunto de prototipos.

En el reconocimiento de patrones, el algoritmo k-nn es usado como método de clasificación de objetos (elementos) basado en un entrenamiento mediante ejemplos cercanos en el espacio de los elementos k-nn es un tipo de aprendizaje vago (*lazylearning*), donde la función se aproxima solo localmente y todo el cómputo es diferido a la clasificación.

4. Propuesta

Para optimizar el sistema de recomendación con un filtro colaborativo orientado al usuario nos enfocamos en el procesamiento de los datos para aumentar la velocidad de las consultas para ellos utilizamos la paralelización para el calculo del algoritmo K-NN con Openmp, obteniendo una considerable reducción de tiempo al momento de analizar los *ratings*.

4.1. Arquitectura

Nuestro sistema esta desarrollado en un entorno C++, la estructura para el procesamiento de los datos es un vector de maps, aqui es donde se carga los ratings de los usuarios para cada pelicula/libro. Para reducir el acceso a memoria indexamos el *dataset* "BX-Books.[en] un gestor de bases de datos (Mysql) aun que una base de datos no relacional como MongoDB o Kasandra es mas convencional para este tipo de sistemas utilizamos las consultas a la base de datos para obtener de manera mas

rapida los datos de los productos o usuarios recomendados , ya que el proceso de análisis se utilizan únicamente códigos de usuario o de libros.

4.2. DataSet

Se trabajo con las siguientes base de datos:

Movielens. Es una DataSet de diversas dimensiones, que estan llenos de datos de peliculas, usuarios y ratings.

- MovieLens 27M Dataset
- MovieLens 20M Dataset
- MovieLens 1M Dataset
- MovieLens 100K Dataset

BX-Books. Es una DataSet de diversas dimensiones, en donde se encuentra libros, users y ratings.

5. Pruebas

Se realizo pruebas con la base de datos de movielens de 27 millones, el codigo fuente y los resultados se encuentran en la siguiente direccion `https://github.com/edsonlipa/Sistema-de-Recomendacion-MovieLens`