SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN EN LA WEB

TALLER No. 9	SISTEMAS DE RECOMENDACION

Profesor: Jaime Alberto Guzmán Luna

Contenido del taller:

1. Sistemas de recomendación
    - Contenido
    - Colaborativo

Actividad A

Sistemas de recomendación basados en contenido

El archivo Peliculas.xlsx contiene una lista de películas y algunos atributos. La última columna indica la calificación que un usuario ha dado a las películas que ha visto. Las películas que no tienen calificación no han sido vistas por el usuario.

Preparación de los datos

Ejercicio 1
Algunos valores para los atributos de las películas faltan. Consulte los datos faltantes y complete la información en la tabla. Guárdela en la hoja 1 del libro y llámela "**Items**".

Ejercicio 2
Se busca representar de manera vectorial cada ítem, de modo que se puedan aplicar las técnicas matemáticas y estadísticas para la recomendación; sin embargo, algunos atributos son categóricos. Realice las transformaciones necesarias de modo tal que la matriz de películas contenga valores numéricos: cree una columna por cada género posible e indique con 0 y 1 si la película pertenece a dicho género; convierta en binario la columna que indica si es una saga o no; y convierta a minutos la duración. Guárdela en la hoja 2 del libro y llámela "**Datos preparados**"

Perfil de usuario

Ejercicio 3
El siguiente paso es obtener una descripción numérica de los gustos del usuario. Utilizando únicamente las películas que el usuario ha calificado y la matriz de calificación, obtenga la matriz que representa el perfil del usuario. Para ello siga los siguientes pasos:

> - Multiplique el valor de cada columna de una película por la calificación de dicha película. Ejemplo: Si para P1 el año es 2005 y la calificación es 5.0, el valor de año será 10.025; si la película tiene en la columna Drama 1 y en la columna Acción 0, sus nuevos valores serán 5 y 0 respectivamente.
>
> De esta forma se obtienen una matriz que describe los gustos del usuario por cada película de acuerdo con su importancia. Por ejemplo, si le gustó mucho una película de Animación y poco una de Acción, los vectores de dichas películas se alejarán

significativamente (Se le está dando magnitud a cada vector según la relevancia para el usuario).
- Obtenga la fila de la matriz que representa al usuario: Cree una fila al final de la matriz llamada “Perfil de usuario”. Los valores de cada columna son la suma de todos los valores de la columna (Suma vertical)

- Normalice el perfil de usuario: Normalice la descripción del perfil de usuario para obtener una importancia relativa en cada columna. Note que la columna año representa una suma ponderada, por lo que debe normalizar teniendo en cuenta el total de los pesos (Calificación):

Divida la suma de la columna año entre la suma de todas las calificaciones de las películas.

La normalización de las columnas que representan el género se debe realizar teniendo en cuenta que todas representan el mismo dato (Recordar que se dividió la columna en varias). Por lo tanto, divida el valor de cada género entre la suma de todas las columnas de los géneros.

Normalice las demás columnas según corresponda

- El perfil de usuario debe ser una matriz que contiene una única fila y un valor para cada columna de la matriz de datos preparados que representa su preferencia para dicho atributo.

- Guarde el perfil de usuario, así como todos los pasos realizados en este ejercicio para obtenerlo en la hoja 3 y llámela “Perfil de usuario”

Recomendación de ítems

Ejercicio 4
Utilizando el perfil de usuario y los ítems sin calificar en la matriz de Datos preparados, se debe encontrar un valor numérico para cada película que represente la posible preferencia del usuario hacia dicha película, que aún no ha visto.

Recuerde que cada ítem es representado como un vector, al igual que el perfil del usuario, por lo tanto, puede elegir cualquier función de distancia o similitud para encontrar vectores similares. En este caso usaremos la distancia euclidiana.

Antes de realizar los cálculos, es importante normalizar correctamente los datos. Note que las variables de los géneros están en el rango 0-1, sin embargo, el año y duración tienen rangos muchos mayores.

- Normalice tanto el perfil de usuario como las películas sin calificar en el rango 0-1. Para normalizar puede utilizar la siguiente forma:

$$y = \frac{(x - x_{min})(d2 - d1)}{x_{max} - x_{min}} + d1$$

donde:

- x - valor a normalizar;
- $[x_{max}, x_{min}]$ - rango de valor x;
- $[d1, d2]$ - rango al que será reducido el valor de x.

Halle la distancia euclidiana entre cada película sin calificar y el perfil de usuario. Concluya y recomiende los primeros k=5 películas que debería ver el usuario de acuerdo con sus gustos.
- Guarde todo el proceso y la conclusión en la hoja 4 y llámela "Recomendaciones"

Cold start problem

Ejercicio 5
A. Suponiendo que ingrese una nueva película a la base de datos, con la siguiente información:

Nombre	Año	Genero	Duracion	Saga
Intocable	2011	Comedia	1:53h	No

¿Qué tan probable es que se le recomiende la nueva película al usuario? No es necesario realizar el proceso matemático, simplemente observe los valores de cada uno de sus atributos y compárelos respecto al perfil del usuario. ¿Cómo influye el año, duración y saga en la recomendación? ¿Qué pasa con el género?
B. Suponiendo que un nuevo usuario se registra en el sistema. ¿Qué películas le recomendaría, por qué, y cómo haría el proceso?

Responda a ambas preguntas en la hoja 5 y llámela "*Cold start problem*"

Deducción del perfil de usuario

Ejercicio 6
Suponiendo que no se quiere preguntar por una calificación específica a los usuarios, o ellos no la realizan, ¿Cómo obtendría el perfil del usuario que represente sus preferencias? Suponga que tiene el registro de las acciones de un usuario para cada película así: Sin ver, vista completamente, Vista parcialmente (Pausa), Vista parcialmente (Detenida), Vista parcialmente (Otra película iniciada)
Explique detalladamente como llevaría dicha información a un perfil de usuario (Numérico y que coincida con la matriz). Responda en la hoja 6 y llámela "Deducción de perfil"

Actividad B

Sistemas de recomendación basados en contenido

Registrar en el archivo compartido en Drive de Películas la calificación de las películas que ha visto en una nueva fila en una escala de 0 a 10.
https://docs.google.com/spreadsheets/d/1EviKnK-6pE-DYv8KQM67RTWuREniMUiueAbCTlovbBQ/edit?usp=sharing

Una vez registradas las calificaciones de todos los demás, descargue una copia del archivo y trabaje sobre él

SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN EN LA WEB

Ejercicio 1
Utilizando la matriz de calificaciones, halle su similitud con cada uno de los demás usuarios, utilizando una alguna medida de distancia o similitud. Utilice k = 8 para obtener los vecinos más cercanos, resalte los 8 usuarios más similares. Recuerde que, para un par de usuarios, solo puede usar las películas que ambos usuarios calificaron para poder encontrar la similitud.

Guarde los resultados en la hoja 2 “Similitud de usuarios”

Ejercicio 2
Utilizando únicamente los k = 8 vecinos más cercanos:
Multiplique la sub matriz de películas que usted no ha visto con la matriz de similitud de usuarios, para obtener la matriz de calificaciones ponderada de películas que usted no ha visto. Luego haga la sumatoria correspondiente por cada ítem y saque el promedio ponderado con los usuarios que vieron las películas (Revisar bien el ejemplo de las diapositivas).

Obtenga la matriz de recomendación y resalte las 5 películas que son más recomendadas para ver.

Guarde los resultados en la hoja 3 “Recomendaciones”.

Ejercicio 3
En la hoja 4, responda a la pregunta: ¿El sistema realizado es un sistema usuario-usuario o ítem-item? Justifique