UNIVERSIDAD AUSTRAL | INGENIERÍA

Introduccion a Data Mining – 2020

Caso IDM-1

Integrantes:
Ludwing Carreño
Alejandra Reyes
Carina Carranza
Mónica Orduz Valbuena

Considere el siguiente caso:

Una cadena de hoteles que gestiona numerosos establecimientos en varios países registra información acerca de sus huéspedes en los diferentes hoteles. La gerencia desea implementar un **proyecto de Data Mining** a efectos de extraer el máximo provecho de esos datos. Más precisamente, la cadena desea conocer mejor a sus clientes de manera de poder informarlos sobre eventos especiales, promociones especiales, etc., durante su estancia como así también después de esta.

A la llegada de un huésped, se registran sus datos demográficos y se le pide completar un cuestionario cuando se le hace entrega de una credencial que le permite ingresar a distintos lugares recreativos tales como piscinas, bares, etc. Esos lugares son gratuitos, pero vía la credencial es posible registrar cuando un huésped hace uso de alguno de esos servicios. Además, la credencial sirve como tarjeta de crédito para pagar ciertas bebidas como así también para pagar productos en las pequeñas tiendas que la cadena posee. Todas esas transacciones son registradas en una base de datos.

Por medio de su canal de TV privado, el hotel informa a sus huéspedes sobre los próximos eventos, actividades y promociones especiales, etc. Sin embargo, puesto que los huéspedes pasan la mayor parte de su tiempo afuera y no mirando TV, el hotel necesita un sistema para enviar anuncios altamente personalizados de manera de garantizar que sólo se envíen anuncios en los que el huésped esté posiblemente interesado. Además, el sistema de data mining también tiene que dar una sugerencia razonable sobre qué anuncios enviar a un huésped particular ya durante los primeros días de su estadía, cuando el sistema tiene pocas posibilidades de aprender los hábitos de ese huésped particular. Después de la estadía, se le solicita al huésped que complete un formulario de evaluación. Este formulario contiene una listade preguntas en las que se debe consignar un puntaje de 1 a 5. Para cada respuesta se pueden explicar los motivos o agregar comentarios adicionales breves.

Durante el invierno, la cadena envía publicaciones de una selección de sus hoteles a antiguos huéspedes, para obtener nuevas reservas en uno de sus hoteles. La selección para este mailing personalizado tiene que ser hecha de modo tal que sólo las publicaciones de los hoteles más interesantes para un huésped particular son enviadas a ese huésped. Por lo expresado, es importante considerar que en la mayoría de los casos un huésped de hotel no elige muchas veces exactamente el mismo hotel.

En conclusión, el sistema de data mining debe proporcionar información para:

1. **Decidir cuáles anuncios de TV van a ser enviados a los huéspedes durante su estadía; un subproblema de este primer problema es decidir cuales anuncios van a ser enviados durante los primeros días de la estadía de un huésped.**
2. **Decidir sobre la selección de hoteles para el mailing privado durante el verano.**

Objetivo
Establecer patrones sobre el comportamiento y preferencias de los clientes de la cadena, para ofrecer opciones de mercado compatibles y de esta manera realizar una segmentación de mercado competitiva.
Este objetivo se toma a partir de la consideración inicial: "Conocer mejor a sus clientes de manera de poder informarlos sobre eventos especiales, promociones especiales, etc."

Parte A - Aplicación de proceso KDD al escenario de negocio

i. Recolectar la información (data):
Considerar los datos generados a partir del registro y el uso de la credencial:
a) Datos demográficos (Perfil)
b) Acceso a lugares recreativos (Preferencias de espacios y tipo de ocio)
c) Consumo de bienes y servicios (Preferencias de productos)

ii. Selección de los datos
Tomar de ese "Data Warehouse" los microdatos relacionados con el perfil: nombre, edad, sexo, lugar de residencia, datos de contacto (correo, celular), fecha de la estadía, lugar, hotel, días de alojamiento, tipo de habitación, tipo de visita (Ocio o trabajo), si viene acompañado (Familia, compañeros de trabajo), financiamiento (Crédito, débito, financiado por empresa).

De igual manera se selecciona: lugares del hotel que frecuenta, horario en el que accede al lugar, tiempo en cada lugar, productos que consume, quién lo acompaña y método de pago.

iii. Preprocesamiento
Revisar los datos, aplicación de histogramas y boxplot, identificación de variables nulas, NA, datos atípicos, resúmenes generales de las variables, estrategias para eliminar ruidos, normalización de los datos.

iv. Transformación de los datos
Transformar datos para su procesamiento, incluye el discretizar ciertas variables,

determinación de la estrategia a utilizar frente a los outliers, registros vacíos, NA y datos nulos. También se puede considerar la reducción del número de variables y la proyección de los datos en otros espacios, para que resulte más fácil encontrar la posible solución, transponer matrices, dinamizar ciertas variables, etc. Este paso resulta crucial, pues de él, depende el éxito o fracaso del proyecto.

v. Minería de datos: En este paso se incluye se utilizan distintas técnicas para generar las recomendaciones. En la parte de B se describirán las técnicas utilizadas mas frecuentemente en los sistemas de recomendación y en la parte C se analizara como aplicarlas al escenario de negocio planteado.

vi. Evaluar los resultados
Determinación según el modelado de datos, la segmentación de clientes, el número de clusters obtenidos y la definición de los patrones en principio obtenidos, como fase de entrenamiento del sistema.

vii. Conocimiento
Revisar el comportamiento de los datos y las desviaciones entre los resultados obtenidos del modelo inicial con los resultados nuevos que se obtienen a través de nueva data.

viii. Evaluar el conocimiento en busca de mejoras
Frente a esa última acción se propende por el desarrollo de acciones de mejora del modelo de entrenamiento, mejoras funcionales, eficiencia y precisión, así como los ajustes sobre la parametrización de las variables y los demás cálculos que tienen lugar.
Formalizamos de esta manera el algoritmo base para poder responder a las necesidades iniciales del cliente y poder ejecutar como tal los planes de mercadeo, con probabilidades de éxito que superen el 50%.

Parte B - Data Mining - Evaluación de los distintos tipos de Sistemas de Recomendación

- El objetivo de un sistema de recomendación es realizar recomendaciones relevantes para los usuarios. La opción más simple seria sugerir siempre aquellas actividades mas populares. Pero para lograr mejores resultados debemos recurrir a alguno de los métodos utilizados con mayor frecuencia en los sistemas recomendación, los cuales pueden dividirse en: collaborative filtering methods o content based method.

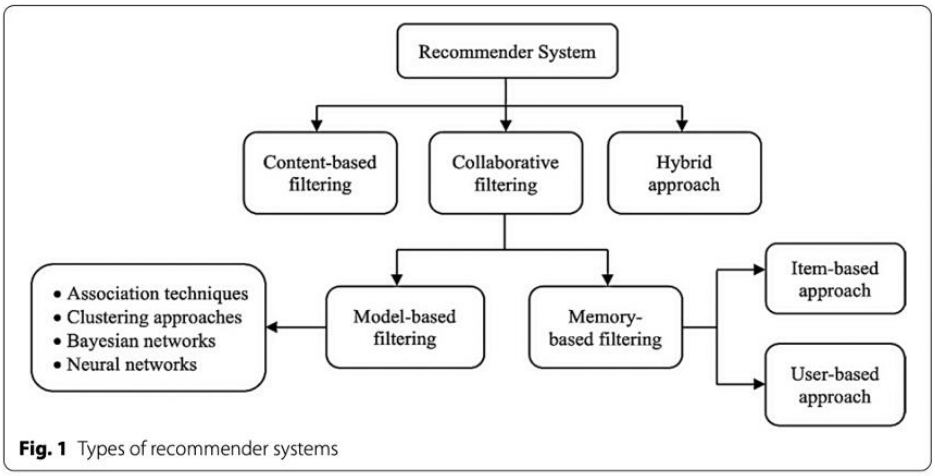
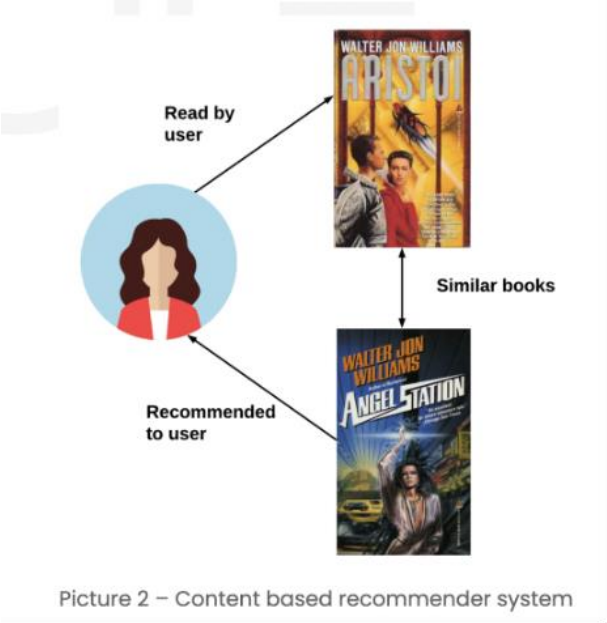


Fig. 1 Types of recommender systems

Content based Method

- Este método para obtener las recomendaciones puede enfocarse como un problema de clasificación (predecir si a un usuario le gusta o no un item) o como un problema de regresión (predecir el rating que le dará un usuario a un item). En ambos casos, se utilizara un modelo basado en las características del usuario y/o del item a nuestra disposición.

- Como desventaja, estos métodos ofrecen poca diversidad en los resultados (efecto burbuja, donde el usuario queda circunscripto a una burbuja de previos intereses).Se precisa recolectar muchos datos de los usuarios y sus preferencias para obtener una buena recomendación con este método.



Picture 2 – Content based recommender system

Collaborative filtering methods

- Se basan en las interacciones previas de los usuarios las cuales pueden guardarse en una matriz de user-item interaction.

A sí mismo, este método se subdivide en dos categorías:

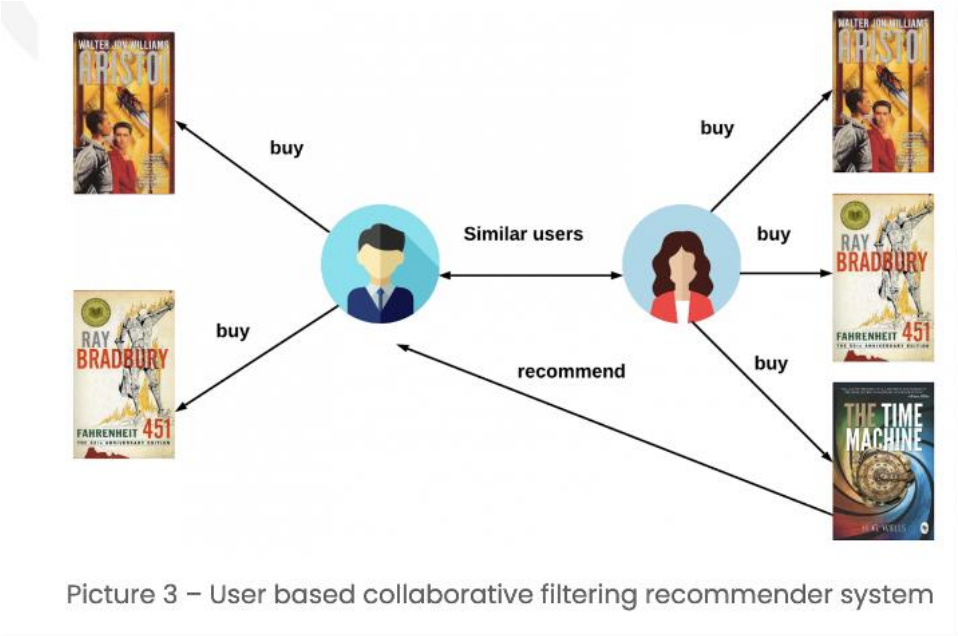
  - **Memory-based:** utiliza solamente las interacciones registradas, no se asume ningún modelo latente.
  - **Model-based**: asume que existe un modelo que explica las interacciones user-items y trata de descubrirlo a fin poder realizar predicciones.

Entre las características de los collaborative filtering approach están:

  - Fácil implementación.
  - A medida que mas usuarios lo utilizan se obtienen mejores recomendaciones.
  - Presentan 'cold start problem' en mayor medida que el content based approach. Se dificulta la tarea de realizar recomendaciones para un nuevo usuario ya que no se cuenta con una matriz de interacciones del mismo.

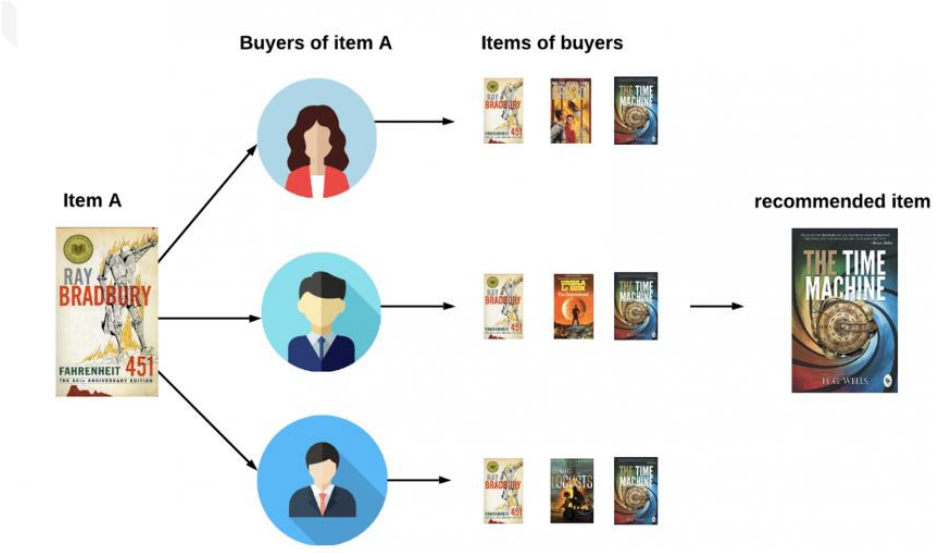
Dentro de los **Memory-based** methods se puede hacer otra subdivisión:

- **User-User Method:** trata de identificar usuarios que tengan un perfil de interacciones similar al del usuario que se está analizando , para luego sugerir ítems que sean populares para ese tipo de usuarios y que sea un item nuevo para el usuario en cuestión. Podria resumirse como “users that are similar to you also liked”.



Picture 3 – User based collaborative filtering recommender system

- **Item-Item Method: t**rata de identificar items similares a aquellos con los cuales el usuario ya ha interactuado positivamente. Podria resumirse como “users who liked this item also liked”.



Picture 4 – Item based collaborative filtering recommender system

Métodos híbridos

- Se utilizan varios métodos asumiendo que un método mixto dará mejores resultados que un solo método.

Parte C - Resumen del Diseño Propuesto y Conclusiones

En base a los métodos descriptos anteriormente, proponemos utilizar un **modelo mixto:**

- Se utilizará un **método model based al comienzo de la estadía**, tomando la información demográfica obtenida en la encuesta de ingreso se hará un clustering con usuarios que tengan datos demográficos similares. De esta manera, se intenta subsanar el problema del cold start , y se ofrecen opciones durante los primeros días de estadía del huésped.
    - El **Clustering** Jerárquico, ofrece la posibilidad de tener la representación de los grupos en árboles y a su vez, no tienen que necesariamente pre definirse el número de agrupaciones, algo sensato para este caso, al menos de forma inicial. A partir del uso de dendrogramas, se pueden establecer preferencias, consumos y capacidad adquisitiva, estableciendo los segmentos de mercado, lo que responde a la posibilidad de utilizar publicidad acorde para cada grupo. De igual manera, se puede llegar a establecer horarios, respondiendo al siguiente subproblema o necesidad. Con el uso de los clusters, se determina los horarios y días en los que resulta más probable atacar a través de las estrategias de mercado a los clientes objetivo.

- Luego, a medida que **el usuario comience a realizar actividades** puede comenzar a utilizarse un modelo colaborativo que permita ajustar las sugerencias de acuerdo a los ratings que va obteniendo en sus interacciones.
- Dentro del **aproach colaborativo** utilizamos el **método user-user** para lo cual:

    1. Se crea una matriz de user -item -> en la cual se van incluyendo las actividades que el usuario empieza a realizar a largo de su estadía.
    2. Se crea una matriz de similtud user-user -> en esta matriz se calcula un coeficiente de similitud entre los usuarios.
    3. Se busca en la matriz user-user , otros usuarios que hayan realizado las mismas actividades que el usuario ha realizado hasta el momento
    4. Luego cuando se encuentra un huésped similar se busca en la matriz user-item que otras actividades ha realizado. Esta actividad similar será recomendada a nuestro nuevo huésped.

- Finalmente, para el **mailing posterior** a la estadía en el hotel se utilizará un **método colaborativo item-item:**

    1. Se crea una matriz de user -item -> en la cual se pone las actividades que el usuario realizo en su estadía y en caso que haya completado se incluirá también la información de la encuesta de salida (se considera que dicha encuesta tiene un bajo índice de respuesta y no todos los usuarios la completan).
    2. Se crea una matriz de similitud item-item (en este caso hoteles)
    3. Se buscan en la matriz hoteles que tengan actividades similares a los seleccionados por el usuario.
    4. En base a eso se genera un hotel candidato para recomendación.
    5. Filtro del hotel candidato-> se verifica que el hotel candidato no haya sido visitado ya por el usuario. En caso de que ya lo haya visitado, se lo elimina y se

repiten los pasos 3 en adelante.

- Según el momento de la estadía del huésped se aplican distintas técnicas, buscando de esta manera obtener un mejor resultado y subsanar la falta de datos en la instancia inicial. Un modelo híbrido nos permite lograr mejores recomendaciones.

Referencias

- Gil, C. (2018) Métodos de Clustering. Recuperado el 23 de noviembre de 2022 de: https://rpubs.com/Cristina_Gil/Clustering
- https://thingsolver.com/introduction-to-recommender-systems/
- Deepjyoti Roy and Mala Dutta, A systematic review and research perspective on recommender systems – Journal of Big Data.