UNIVERSIDAD AUSTRAL | INGENIERÍA

Introducción a Data Mining

Caso IDM-1

Damian Quiroga
DNI 28.816.690
qdamian@gmail.com

Visión general

El proceso de KDD nos permite procesar los datos disponibles para descubrir patrones que permitan a la cadena de hoteles mejorar sus anuncios de servicios y establecimientos para sus clientes.

1. Limpieza de datos

Tanto los datos ingresados por los huéspedes al llenar formularios o registrado por los establecimientos deben validarse por errores al ingreso, omisiones, errores de tipeo en la carga, etc. Se pueden aplicar distintas técnicas según el tipo de datos, por ejemplo:

- **Detección de errores:**
  - Validación de formato: por ej. para los números de identificación (DNI, pasaporte) y direcciones de correo electrónicos.
  - Mediante outliers: en los datos numéricos como edad, pueden identificarse valores que sean erróneos.
  - Validación de rango: por ej. en los cuestionarios de evaluación. Las respuestas que no estén en el rango esperado (1 a 5) no deberían tenerse en cuenta
  - Detección de inconsistencias: como un servicio o compra efectuada por un huésped en un periodo que no corresponde con su estadía
- **Estandarización:**
  - para los nombres y apellidos guardarlos en un orden consistente
  - para los números telefónicos, de manera de siempre incluir código internacional, código de área y número local de 7 dígitos
  - para los emails siempre convertirlos a minúsculas

Otro punto a tener en cuenta, más allá de la limpieza de los datos, es que los mismos estén capturando información real y actualizada. En particular para los correos electrónicos es necesario validar que las direcciones sean activamente usadas por los huéspedes ya que se desea enviar emails promocionales.

2. Integración de datos

El hecho de tener numerosos establecimientos en distintos países presupone una gran cantidad de datos disponibles, y cierta heterogeneidad en los mismos. Los servicios que el hotel ofrece en un país podrían no estar disponible en otros. Los precios pueden estar expresados en distintas monedas. El idioma podría ser diferente y también el formato en el que se describe la información, como direcciones y códigos postales.

Para obtener el máximo provecho de la información disponible, conviene integrar la información de distintos establecimientos. Se podría utilizar una estrategia de integrar los datos por país, por región o globalmente. Se deben tener en cuenta también los aspectos legales en caso de procesar datos, según la legislación vigente en cada país.

Una vez definido esto se requiere implementar los procesos que toman los datos, los procesan y cargan en un Data Warehouse para su posterior utilización.

3. Selección de datos

Identificamos características útiles para representar los datos según el caso:

Objetivo de negocio	Datos a utilizar
Decidir cuáles anuncios de TV van a ser enviados durante los primeros días de la estadía de un huésped	• Datos demográficos • Cuestionario de check-in • Servicios ofrecidos por cada establecimiento
Decidir cuáles anuncios de TV van a ser enviados a los huéspedes durante el resto de su estadía	Los anteriores más: • Servicios utilizados • Compras realizadas
Decidir sobre la selección de hoteles para el mailing privado.	Los anteriores más: • Cuestionario de check-out • Ubicación de los establecimientos

4. Transformación de datos

Para cada servicio o producto que puede adquirirse en el hotel sería de utilidad para el posterior minado de datos poder consultar qué atributos tienen los huéspedes que los consumen, tales como edad, sexo, nacionalidad, ocupacion, tipo de habitacion, tiempo de permanencia, etc. y estos datos deberian poder consultarse con distintos tipos de filtrado (por ej. por rango de tiempo o costo del servicio) y tipo de agregación (por ej. para un hotel específico o para todos los hoteles de una región geográfica).

Estas relaciones entre producto/servicio y los datos de los consumidores y datos históricos podría realizarse con relativa facilidad si los huéspedes accedieron a los mismos a través de la credencial del hotel. En caso contrario, podría derivarse en algunos casos a través del número de habitación o tarjeta de crédito del consumidor.

En cualquier caso es necesario almacenar los datos en una base de datos que permita hacer estas consultas. Se necesita por lo tanto algún proceso que permita ir actualizando esta base de datos en la medida que se registran nuevos consumos.

5. Minado de datos

Se podrían emplear distintas técnicas de data mining con el objetivo de extraer patrones a partir de estos datos.

1. Minado de reglas de asociación

Considerando los datos disponibles al momento de registrarse, un sistema de minado de datos podría generar reglas de asociación del tipo:

Regla 1: edad(X, “40..59”) ∧ habitación(X, “premium, suite”) ⇒ utiliza(X, “spa”) [soporte = 2%, confianza = 60%]

En este ejemplo, la regla indica que de los huéspedes bajo estudio, 2% tienen edad entre 40 y 59 años y han alquilado un tipo de habitación premium o suite, y han utilizado el servicio de spa. Hay una probabilidad de 60% de que un huésped con este rango etario y tipo de habitación utilice el servicio de spa.

Para la selección de hoteles para el mail privado, una característica a tener en cuenta es la ubicación del hotel visitado y otros establecimientos que ofrezcan servicios similares a aquellos que son de interés para el huesped.

2. Agrupación (clustering)

Se podrían segmentar a los huéspedes en categorías/grupos bajo el supuesto de que un huésped que comparte las características de un grupo tiene mayor probabilidad de compartir sus intereses. Es decir, una vez decidido a qué grupo corresponde un huésped, se elegirían para la publicidad por el canal privado de TV los servicios que sean más consumidos o que tengan mejor puntuación en las encuestas de evaluación, utilizando la información histórica de los huéspedes asociados a esa categoría.

Si los datos que se utilizan para realizar esta segmentación se limitan a aquellos disponibles al momento de check-in, o sea los datos demográficos y la encuesta inicial, la misma podría utilizarse para obtener las sugerencias para la publicidad desde el check in.

6. Evaluación de patrones

Debemos identificar qué patrones son verdaderamente interesantes con el objetivo de publicitar los establecimientos y servicios de la cadena de hoteles. La evaluación depende del tipo de técnica utilizada. En nuestro caso:

1. Minado de reglas de asociación

Utilizando los datos recolectados, para cada regla calculamos el soporte y confianza. Por ej. para la Regla 1 planteada anteriormente,

X = edad(X, “40..59”) ∧ habitación(X, “premium, suite”)
Y = utiliza(X, “spa”)

Calculamos:

- Soporte, ¿qué porcentaje de uso de servicios satisfacen una regla de asociación?
  - soporte(X ⇒ Y) = P(X ∪ Y),
- Confianza, ¿qué grado de certeza hay en las asociaciones detectadas?
  - confianza(X ⇒ Y) = P(Y |X)

2. Agrupación (clustering)

Según el algoritmo de clustering y sus parámetros podemos obtener distintos agrupamientos. Podemos evaluar la calidad de estos grupos en base a su capacidad de predecir los intereses de los clientes en base a su categoría. Si consideramos el caso de un huésped utilizando un servicio que es el más consumido por su categoría como un Verdadero Positivo (TP) y el caso contrario como un Falso Positivo (FP), podemos calcular la precisión en las predicciones como TP / (TP + FP).

7. Presentacion de conocimiento

Se utilizan distintas representaciones para facilitar la visualización e interpretación del conocimiento minado. Se pueden presentar a través de Dashboards buscando un balance entre flexibilidad y facilidad de uso por parte de la gerencia.

Para poder medir el impacto que tiene el proyecto de data mining es necesario poder comparar

Información deseada	Visualización
¿Qué recomendaciones surgen del análisis de data mining?	- servicios recomendados para el sistema de TV interno con la información disponible al momento de check-in: cantidad de anuncios vs. tipo de servicio - hoteles recomendados para la publicidad vía email: cantidad de menciones en los emails vs. establecimiento de hotel.
¿Qué tan relevantes y efectivas son las recomendaciones de servicios?	- cantidad de recomendaciones de un servicio vs. uso del mismo - % de uso de cada servicio por periodo de tiempo - servicio más recomendado según el tipo de huéspedes
¿Las recomendaciones mejoran la satisfacción del cliente?	- promedio de satisfacción (según encuestas) por periodo de tiempo - promedio de satisfacción (según encuestas) por servicio

¿Qué tan relevantes y efectivas son las recomendaciones de hoteles vía email?	- tasa de efectividad de las recomendaciones: reservaciones de un establecimiento en relación al número de recomendaciones - % de reservaciones vs. canal por el cual el huésped conoció la posibilidad, según encuestas - % de emails que llegan a potenciales clientes vs. retornados por estar la dirección desactualizada - % de emails que resultan en consultas, a través del link provisto en el email
¿Se adaptan las recomendaciones al perfil e historial de consumos de cada huésped?	- efectividad de los anuncios tras ´n´ dias de estadia vs. la efectividad de anuncios en el dia 1