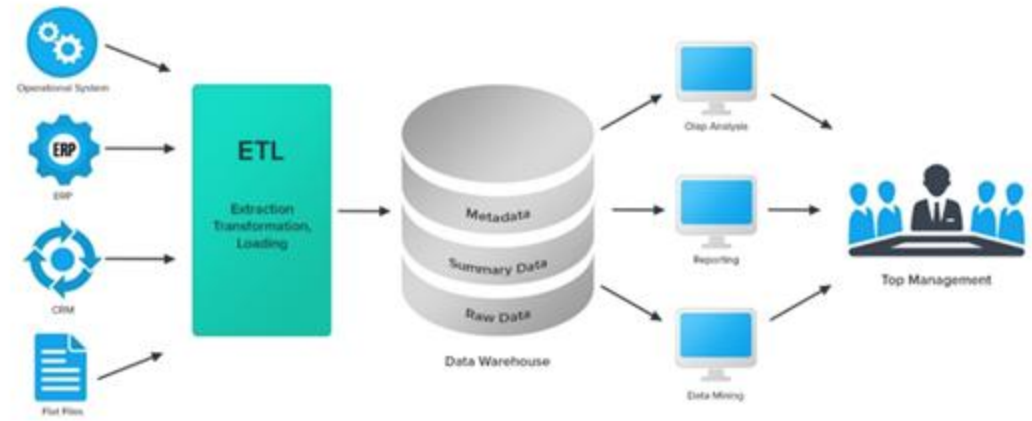
FACULTAD DE INGENIERÍA
ESCUELA DE COMPUTACIÓN



Datawarehouse y Minería de Datos

Guía #9: Reglas de Asociación

Minería de Datos o Data Mining

Es un **conjunto de técnicas** a las que se le aplica la tecnología, con el fin de procesar, mediante exploración, una inmensidad de volúmenes de datos, que de manera automatizada o parcialmente automatizada hace posible ***localizar patrones, tendencias o incluso dar respuestas futuribles en escenarios reales o imaginables*** en el entorno de la empresa y en un determinado contexto, ***siendo capaz de convertir los datos en información y la información en conocimiento***, para así poder optimizar las decisiones empresariales.

Ejemplo 1 - Análisis de créditos bancarios

El primer ejemplo pertenece al ámbito de la banca. Un banco por Internet desea obtener reglas para predecir las personas de las que solicitan un crédito no lo devuelven. La entidad bancaria cuenta con los datos correspondientes a los créditos concedidos con anterioridad a sus clientes (cuantía del crédito, duración en años...) y otros datos personales como el salario del cliente, si posee casa propia, etc. Algunos registros de clientes de esta base de datos se muestran en la siguiente tabla.

IDC	D-crédito (años)	C-crédito (euros)	Salario (euros)	Casa propia	Cuentas morosas	...	Devuelve-crédito
101	15	60.000	2.200	sí	2	...	no
102	2	30.000	3.500	sí	0	...	sí
103	9	9.000	1.700	sí	1	...	no
104	15	18.000	1.900	no	0	...	sí
105	10	24.000	2.100	no	0	...	no
...	...	...	...	...	...	...	...

Ejemplo 1 - Análisis de créditos bancarios

A partir de algunas reglas, las técnicas de minera de datos podrían sintetizar reglas, como por ejemplo:

SI Cuentas-Morosas > 0 **ENTONCES**

Devuelve-crédito = no

SI Cuentas-Morosas = 0 Y [(Salario > 2.500) Ó (D-crédito > 10)] **ENTONCES**

Devuelve-crédito = sí

El banco podría entonces utilizar estas reglas para determinar las acciones a realizar en el trámite de los créditos: si se concede o no el crédito solicitado, si es necesario pedir avales especiales, etc.

Ejemplo 2 - Análisis de la cesta de compra

Un supermercado quiere obtener información sobre el comportamiento de compra de sus clientes. Piensa que de esta forma puede mejorar el servicio que les ofrece: reubicación de los productos que se suelen comprar juntos, localizar el emplazamiento idóneo para nuevos productos, etc. Para ello dispone de la información de los productos que se adquieren en cada una de las compras o cestas.

Idcesta	Huevos	Aceite	Pañales	Vino	Leche	Mantequilla	Salmón	Lechugas	...
1	sí	no	no	sí	no	sí	sí	sí	...
2	no	sí	no	no	sí	no	no	sí	...
3	no	no	sí	no	sí	no	no	no	...
4	no	sí	sí	no	sí	no	no	no	...
5	sí	sí	no	no	no	sí	no	sí	...
6	sí	no	no	sí	sí	sí	sí	no	...
7	no	no	no	no	no	no	no	no	...
8	sí	sí	sí	sí	sí	sí	sí	no	...
...	...	...	...	...	...	...	...	...	...

Ejemplo 2 - Análisis de la cesta de compra

Analizando los datos del supermercado podríamos encontrar los siguientes patrones:

- El 100 por ciento de las veces que se compran pañales también se compra leche
- El 50 porciento de las veces que se compran huevos, se compra aceite
- El 33 por ciento de las veces que se compra vino y salmón, entonces se compra lechugas.

También se puede analizar cuáles de estas asociaciones son frecuentes, porque una asociación muy estrecha entre dos productos puede ser poco frecuente y, por tanto, poco útil.

Reglas de Asociación

Describe una relación de asociación entre los elementos de un conjuntos de datos.

Por ejemplo:

- Un supermercado que desea conocer que productos suelen comprarse conjuntamente, y así mejorar la distribución de los productos en estanterías.
- En un servidor web podemos conocer cuáles son los itinerarios más seguidos por los visitantes a las páginas web, y entonces, utilizar esta información para estructurar las páginas web en el servidor.
- Estudiantes que cursa Matemáticas tienden a cursar Física.

Origen de las Reglas de Asociación

Las reglas de asociación *surgieron inicialmente para afrontar el análisis de la cesta de compra de los comercios*. En este contexto, las diferentes cestas de compra se pueden expresar formando una base de datos en una sola tabla.

Las filas de esta tabla se refieren a una cesta de un supermercado, mientras que las columnas son cada uno de los productos en venta del supermercado.

La tabla solo contiene valores. Un 1 en la posición (*i, j*) indica que las cesta *i* incorpora el producto *j*, mientras que un 0 indica que el cliente no ha adquirido el producto.

Origen de las Reglas de Asociación

La siguiente tabla podría ser un ejemplo de una base de datos de este tipo.

	Vino "El cabezón"	Gaseosa "Chispa"	Vino "Tío Paco"	Horchata "Xufer"	Bizcochos "Goloso"	Galletas "Trigo"	Chocolate "La vaca"
T1	1	1	0	0	0	1	0
T2	0	1	1	0	0	0	0
T3	0	0	0	1	1	1	0
T4	1	1	0	1	1	1	1
T5	0	0	0	0	0	1	0
T6	1	0	0	0	0	1	1
T7	0	1	1	1	1	0	0
T8	0	0	0	1	1	1	1
T9	1	1	0	0	1	0	1
T10	0	1	0	0	1	0	0

Una regla de asociación es una proposición probabilística sobre la ocurrencia de ciertos estados en una base de datos. Una típica regla de asociación sería:

SI bizcochos "Goloso" **Y** horchata "Xufer" **ENTONCES** galletas "Trigo"

Definición de Reglas de Asociación

- Forma general: *X* → *Y*, donde *X* e *Y* son conjuntos de ítems.
- *X* es denominado el antecedente de la regla e *Y* su consecuente.

Dada la regla de asociación, se suele trabajar con dos medidas para conocer la calidad de la regla: **cobertura** (support) y **confianza** (confidence).

- ***La cobertura*** (también denominada soporte) de una regla se define como el número de instancias que la regla predice correctamente (también se utiliza el porcentaje).
- Por otra parte, ***la confianza*** (también conocida como precisión) mide el porcentaje de veces que la regla se cumple cuando se puede aplicar.
- Mejorando la confianza, **Lift** este indicador proporciona soporte al conjunto de datos. Si el valor de Lift=1 indica que el conjunto aparece una cantidad de veces acorde a lo esperado, si Lift>1 indica que ese conjunto aparece una cantidad de veces superior a lo esperado y si el Lift<1 indica que ese conjunto aparece una cantidad de veces inferior a lo esperado.

Definición de Reglas de Asociación

Soporte: Frecuencia relativa de una regla sobre el total de transacciones

Ejemplo: {Leche, Pañales} → {Cerveza}

$$s = \frac{\sigma(X)}{|T|}$$

Transacciones

T	Items
1	pan, leche
2	pan, pañales, cerveza, huevos
3	leche, pañales, cerveza, diario
4	pan, leche, pañales, cerveza
5	pan, leche, pañales, diario

$$s = \frac{\sigma(\{Leche, Pañales, Cerveza\})}{|T|}$$

Transacciones

T	Items
1	pan, leche
2	pan, pañales, cerveza, huevos
3	leche, pañales, cerveza, diario
4	pan, leche, pañales, cerveza
5	pan, leche, pañales, diario

$$s = \frac{2}{5} = 0.4$$

EL 40% de transacciones compraron leche, pañales y cervezas.

Definición de Reglas de Asociación

Confianza: Mide que tan confiable es la suposición hecha por la regla

Ejemplo: {Leche, Pañales} → {Cerveza}

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Transacciones

T	Items
1	pan, leche
2	pan, pañales, cerveza, huevos
3	leche, pañales, cerveza, diario
4	pan, leche, pañales, cerveza
5	pan, leche, pañales, diario

$$s = \frac{\sigma(\{Leche, Pañales, Cerveza\})}{|T|}$$

Transacciones

T	Items
1	pan, leche
2	pan, pañales, cerveza, huevos
3	leche, pañales, cerveza, diario
4	pan, leche, pañales, cerveza
5	pan, leche, pañales, diario

$$c = \frac{2}{3} = 0.67$$

Un total de 2 ocurrencias de los tres artículos: leche, pañales y cerveza.

Numero de filas que contiene leche y pañales, ya que es un subconjunto de leche, pañales y cerveza.

Definición de Reglas de Asociación

Lift: Confianza a la regla dividido por el soporte del consecuente

Ejemplo: {Leche, Pañales} → {Cerveza}

$$Lift = \frac{c(X \rightarrow Y)}{s(Y)}$$

Transacciones

T	Items
1	pan, leche
2	pan, pañales, cerveza, huevos
3	leche, pañales, cerveza, diario
4	pan, leche, pañales, cerveza
5	pan, leche, pañales, diario

$$Lift = \frac{c(\{Leche, Pañales\} \rightarrow \{Cerveza\})}{s(\{Cerveza\})}$$

$$c(\{Leche, Pañales\} \rightarrow \{Cerveza\}) = 0.67$$

$$s(\{Cerveza\}) = \frac{3}{5} = 0.6$$

$$Lift = \frac{0.67}{0.6} \approx 1.117$$

Lift > 1: podemos tener certeza que la probabilidad de nuestra regla aumento cuando se compra leche y pañales.

Lift = 1: La probabilidad no se ve afectada.

Lift < 1: La probabilidad es más baja.

Algoritmos Apriori y FP-Growth

Algoritmo Apriori

Propuesto por Agrawal, es uno de los primeros y más populares algoritmos para la minería de reglas de asociación. Este algoritmo descubre todas reglas de asociación en dos fases, ***usando como parámetros un valor de soporte y confianza*** mínimos. Difiere de algoritmos previos en la manera en que los conjuntos de elementos son considerados frecuentes y el mecanismo por el cual son generados, obteniendo así un mejor rendimiento en el orden de magnitud para un conjunto de datos grande.

Algoritmo FP Growth

Basado en una mejora del algoritmo Apriori propuesto por Han, define una primera fase que consiste en descubrir los elementos frecuentes, y en una segunda fase en la que se generan las reglas de asociación de los elementos frecuentes encontrados, basado a parámetros de soporte y confianza mínimos. La principal diferencia con el algoritmo Apriori es la implementación usada en FP Growth, la cual es más ***eficiente al hacer uso de un árbol de elementos frecuentes*** que puede ser procesado más rápidamente que la estructura de datos usada en Apriori.

Comencemos con nuestra práctica