Herramientas de R para la investigación de mercados

3. Market Basket Analysis & Association Rules (Análisis de la cesta de la compra & Reglas de asociación)

José Ignacio Casas

The Incredible Story Of How Target Exposed A Teen Girl's Pregnancy

Gus Lubin Feb 16, 2012, 4:27 PM

Target broke through to a new level of customer tracking with the help of statistical genius Andrew Pole, according to a New York Times Magazine cover story by Charles Duhigg.

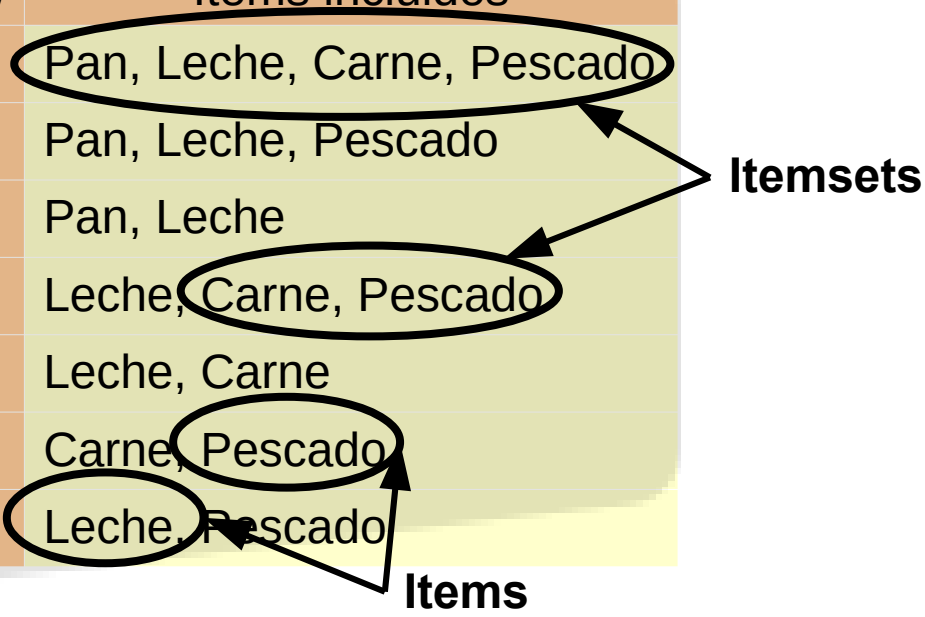
Pole identified 25 products that when purchased together indicate a women is likely pregnant. The value of this information was that Target could send coupons to the pregnant woman at an expensive and habit-forming period of her

Agenda

1. Transacciones & Items
2. Itemsets frecuentes con ***Apriori***
3. Itemsets frecuentes con ***Eclat***
4. Reglas de asociación
5. Evaluación de las Reglas
6. MBA & AR con {arules}

1. Transacciones & Items (1)

Transacciones (cestas de la compra)	Items incluidos
id1	Pan, Leche, Carne, Pescado
id2	Pan, Leche, Pescado
id3	Pan, Leche
id4	Leche, Carne, Pescado
id5	Leche, Carne
id6	Carne, Pescado
id7	Leche, Pescado



1. Transacciones & Items (2)

- **Soporte del item o itemset X**: número de transacciones que contienen X dividido por el total de transacciones

$$soporte(X) = \frac{frec(X)}{N} = P(X)$$

- **Confianza ("Si X entonces Y")**: probabilidad de que una transacción que contiene los items de X, también contenga los items de Y

$$confianza(X \Rightarrow Y) = \frac{soporte(X \cup Y)}{soporte(X)}$$

2. Itemsets frecuentes con ***Apriori*** (1)

- **Itemsets frecuentes = itemsets** con una frecuencia mayor o igual a un determinado soporte mínimo
- Algoritmo ***Apriori***:
  - búsqueda por niveles de complejidad: de menor a mayor tamaño de **itemsets**
  - norma: si un **itemset** no es frecuente, ningún **itemset** de mayor tamaño que contengan al primero puede ser frecuente
  - se inicia identificando los **items** individuales que aparecen en el total de transacciones con un soporte por encima de un mínimo establecido
  - se extienden los candidatos añadiendo un nuevo **item** y se eliminan aquellos que contienen un subconjunto infrecuente o que no alcanzan el soporte mínimo
  - este proceso se repite hasta que el algoritmo no encuentra más ampliaciones exitosas de los **itemsets** previos o cuando se alcanza un tamaño máximo

2. Itemsets frecuentes con ***Apriori*** (2)

En nuestro ejemplo consideraremos i**temsets frecuentes** aquéllos que aparecen al menos en tres transacciones: soporte mínimo = 3/7 = 0,43

Itemset (k = 1)	Soporte
{Carne}	0,57
{Leche}	0,86
{Pan}	0,43
{Pescado}	0,71

Itemset (k = 2)	Soporte
{Carne, Leche}	0,43
~~{Carne, Pan}~~	~~0,14~~
{Carne, Pescado}	0,43
{Leche, Pan}	0,43
{Leche, Pescado}	0,57
~~{Pan, Pescado}~~	~~0,29~~

Itemset (k = 3)	Soporte
~~{Carne, Leche, Pan}~~	Tiene conjunto infrecuente
~~{Carne, Leche, Pescado}~~	~~0,29~~
~~{Carne, Pan, Pescado}~~	Tiene conjunto infrecuente
~~{Leche, Pan, Pescado}~~	Tiene conjunto infrecuente

3. Itemsets frecuentes con ***Eclat*** (1)

El algoritmo ***Eclat*** (**E**quivalence **Cla**ss **T**ransformation) analiza, en primer lugar, las transacciones en las que aparece cada itemset de k = 1:

Itemset (k = 1)	Transacciones	Soporte
{Carne}	id1, id4, id5, id6	0,57
{Leche}	id1, id2, id3, id4, id5, id7	0,86
{Pan}	id1, id2, id3	0,43
{Pescado}	id1, id2, id4, id6, id7	0,71

3. Itemsets frecuentes con ***Eclat*** (2)

Calculando todas las posibles intersecciones de la columna Transacciones de la tabla anterior obtenemos los itemsets de longitud k = 2, teniendo en cuenta un soporte mínimo = 3/7 = 0,43:

Itemset (k = 2)	Transacciones	Soporte
{Carne, Leche}	id1, id4, id5	0,43
~~{Carne, Pan}~~	~~id1~~	~~0,14~~
{Carne, Pescado}	id1, id4, id6	0,43
{Leche, Pan}	id1, id2, id3	0,43
{Leche, Pescado}	id1, id2, id4, id7	0,57
~~{Pan, Pescado}~~	~~id1, id2~~	~~0,29~~

3. Itemsets frecuentes con ***Eclat*** (3)

Con las intersecciones de las Transacciones de la tabla anterior obtenemos los itemsets de longitud k = 3, teniendo en cuenta un soporte mínimo = 3/7 = 0,43:

Itemset (k = 3)	Transacciones	Soporte
~~{Carne, Leche, Pescado}~~	~~id1, id4~~	~~0,29~~

4. Reglas de asociación

Buscaremos únicamente reglas con una **confianza** igual o superior a 0,7, es decir, que la regla se cumpla un 70% de las veces

Reglas	Confianza	Confianza
{Pan} => {Leche}	soporte {Pan, Leche} / soporte {Pan}	0,43 / 0,43 = 1
{Leche} => {Pan}	soporte {Pan, Leche} / soporte {Leche}	0,43 / 0,86 = 0,5
{Leche} => {Carne}	soporte {Leche, Carne} / soporte {Leche}	0,43 / 0,86 = 0,5
{Carne} => {Leche}	soporte {Leche, Carne} / soporte {Carne}	0,43 / 0,57 = 0,75
{Leche} => {Pescado}	soporte {Leche, Pescado} / soporte {Leche}	0,43 / 0,86 = 0,5
{Pescado} => {Leche}	soporte {Leche, Pescado} / soporte {Pescado}	0,43 / 0,57 = 0,75
{Carne} => {Pescado}	soporte {Carne, Pescado} / soporte {Carne}	0,43 / 0,57 = 0,75
{Pescado} => {Carne}	soporte {Carne, Pescado} / soporte {Pescado}	0,43 / 0,71 = 0,6

5. Evaluación de las Reglas

- **Lift**: compara la frecuencia observada de una regla con la frecuencia esperada simplemente por azar (si la regla no existiera). El valor **lift** de una regla “si X, entonces Y” es:

$$lift\left(X \Rightarrow Y\right) = \frac{soporte\left(X \cap Y\right)}{soporte\left(X\right) * soporte\left(Y\right)}$$

- Un valor de **lift** igual a 1 o cercano indica 1 que la regla de asociación es aleatoria
- Una regla con un **lift** de 18 implica que ambos items son 18 veces más probable de ser comprados juntos en comparación de las compras cuando se les supone no relacionados

6. MBA & AR con {arules}