REGLAS DE ASOCIACIÓN

Mg Sc. Meza Rodríguez, Aldo Richard

armeza@ulima.edu.pe

Reglas de asociación

minsup = 2

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

C_1 → 1st scan → L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2 → 2nd scan → C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan → L_3

Itemset	sup
{B, C, E}	2

$$Support = \frac{frq(X,Y)}{N}$$

$$Rule: X \Rightarrow Y \rightarrow Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Mg.sc Aldo Meza Rodriguez

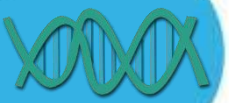
Reglas de asociación

Los algoritmos de reglas de asociación tienen como objetivo encontrar relaciones dentro un conjunto de transacciones, en concreto, items o atributos que tienden a ocurrir de forma conjunta.

Reglas de asociación

OB JETIVO
describir patrones en el comportamiento de las instancias

ALGORITMOS
Apriori, FP-Growth, Eclat



APLICACIONES
Análisis de compras, Análisis de textos, patrones de páginas web
Bioinformática y diagnóstico médico

Reglas de asociación

- Por ejemplo, la siguiente tabla es un ejemplo de una base de datos de transacciones:

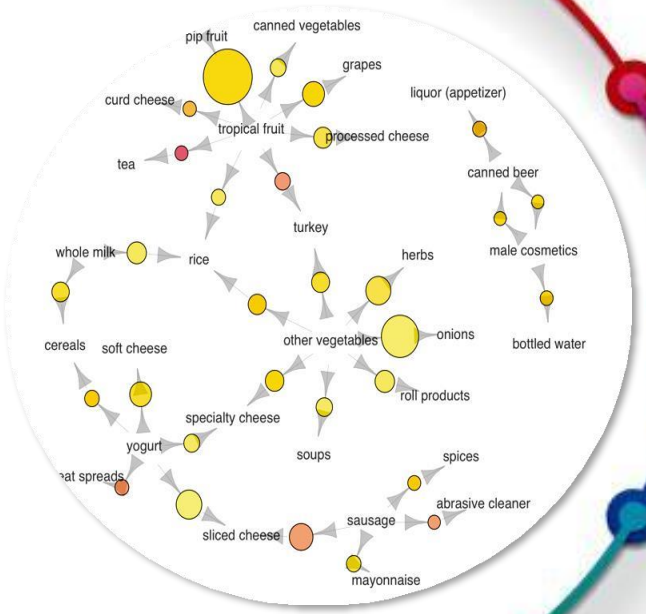
ID	Items
1	Pan, Leche
2	Pan, Pañales, cerveza, huevos
3	Leche, Pañales, Cerveza, Cola
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Cerveza, Cola

$$X \Rightarrow Y$$

$$\{Leche, Pañales\} \Rightarrow Cerveza$$

- Una regla de asociación es un implicancia de la forma $X \Rightarrow Y$, tal que $X \subset I$, $Y \subset I$ y $X \cap Y = \emptyset$.
- X es conocida como la regla antecedente, mientra que Y como la regla consecuente.

Elementos de las reglas de asociación



- ***Items***: artículos que componen una transacción (colección de uno o más ítmes)
- ***Itemset:*** conjunto de *items* de una transacción. Un k-*itemset* es un *itemset* con k artículos. Ej.: productos de la cesta de la compra, páginas web visitadas, etc.
- ***Base de datos transaccional:*** conjunto de transacciones (T= {t1,t2,...,tNt1,t2,...,tN}). Cada transacción viene representada por su conjunto de *items* (I= {i1,i2,...,,...,id})
- ***Soporte:*** Es la frecuencia con la que ocurre un itemset.
- ***Confianza:*** Se define como la fracción de transacciones en las que aparecen los *itemsets* X e Y

Lift: Mide la correlación de dos elementos dada una regla (muestra cómo un conjunto de elementos A afecta al conjunto de elementos B).

Reglas de asociación

$$Soporte(X \Rightarrow Y) = Soporte(XUY) = \frac{cont(XUY)}{N}$$

$$Confianza(X \Rightarrow Y) = \frac{Soporte(XUY)}{Soporte(X)}$$

$$lift(X \Rightarrow Y) = \frac{Soporte(XUY)}{Soporte(X) * Soporte(Y)}$$

ID	Items
1	Pan, Leche,pañales
2	Pan, Pañales, cerveza, huevos
3	Leche, Pañales, Cerveza, Cola
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Cerveza, Cola

Lift = 1, entonces A y B son independientes
Lift > 1, entonces A y B son dependientes entre sí.
Lift < 1, entonces la presencia de A tendrá un efecto negativo en B.

$$X \Rightarrow Y$$

$$\{Leche, Pañales\} \Rightarrow Cerveza$$

$$S(X) = S(Leche, Pañales) = \frac{3}{5} = 0.6$$

$$S(Y) = S(Cerveza) = \frac{4}{5} = 0.8$$

$$S(X \Rightarrow Y) = S(XUY) = S(Leche, Pañales, Cerveza) = \frac{2}{5} = 0.4$$

$$C(X \Rightarrow Y) = P(Y/X) = \frac{S(Leche, Pañales, Cerveza)}{S(Leche, Pañales)} = \frac{2}{3} = 0.67$$

$$lift(X \Rightarrow Y) = \frac{S(Leche, Pañales, Cerveza)}{S(Leche, Pañales) * S(Cerveza)} = \frac{2/5}{(\frac{3}{5}) * (\frac{4}{5})} = 0.83$$

Procedimiento

Para descubrir reglas de asociación, es necesario establecer unos límites mínimos de sorporte y confianza. Así, un itemset frecuente será aquel itemset cuyo soporte supere un mínimo establecido

Pasos

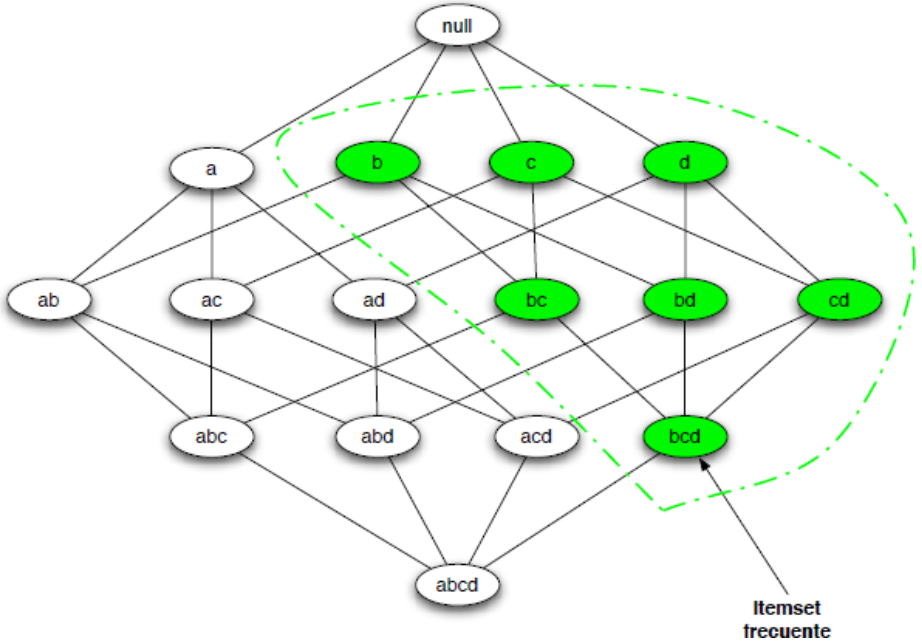
Detectar itemsets frecuentes que superen un soporte mínimo, es decir, cuya ocurrencia supera un mínimo de transacciones

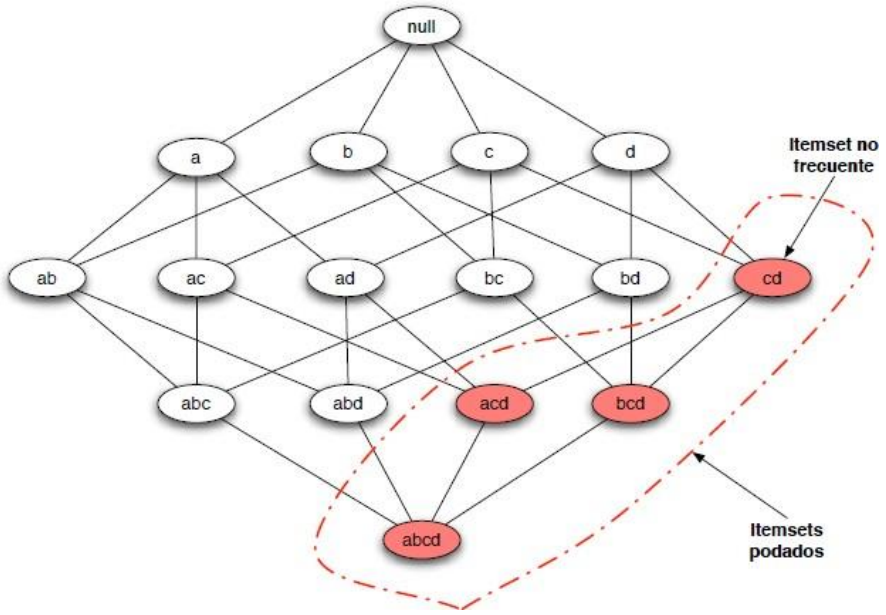
Obtener las reglas de asociación asociadas a dichos itemsets que superen una determinada confianza

El algoritmo Apriori

Apriori fue uno de los primeros algoritmos desarrollados para la búsqueda de reglas de asociación y sigue siendo uno de los más empleados, tiene dos etapas:

- **Identificar** todos los *itemsets* que ocurren con una frecuencia por encima de un determinado límite (*itemsets* frecuentes).
- **Convertir** esos *itemsets* frecuentes en reglas de asociación.





si el itemset {b, c} es frecuente, se puede suponer que {b} es también frecuente (como mínimo aparecerá el mismo número de veces que {b, c}). De igual forma, cualquier adición de items a {b, c} también será frecuente

Si {c, d} es infrecuente, {b, c, d} tampoco será frecuente. Esta propiedad se conoce como **antimonotonicidad**, y permite hacer una poda en el espacio de búsqueda basada en el soporte

Ejemplo del del algoritmo Apriori

Transacción
{A, B, C, D}
{A, B, D}
{A, B}
{B, C, D}
{B, C}
{C, D}
{B, D}

Se procede a identificar los itemsets frecuentes y, a partir de ellos, crear reglas de asociación.

Itemset (k=1)	Ocurrencias	Soporte
{A}	3	0.43
{B}	6	0.86
{C}	4	0.57
{D}	5	0.71

Ej: Considere que un item o itemset es frecuente si aparece en un mínimo de 3 transacciones, es decir, su soporte debe de ser igual o superior a 3/7 = 0.43.

Todos los itemsets de tamaño k = 1 tienen un soporte igual o superior al mínimo establecido, por lo que todos superan la fase de filtrado (poda).

Itemset (k=2)	Ocurrencias	Soporte
{A, B}	3	0.43
{A, C}	1	0.14
{A, D}	2	0.29
{B, C}	3	0.43
{B, D}	4	0.57
{C, D}	3	0.43

A continuación, se generan todos los posibles itemsets de tamaño k = 2 que se pueden crear con los itemsets que han superado el paso anterior y se calcula su soporte.

Itemset (k=2)	Ocurrencias	Soporte
{A, B}	3	0.43
{B, C}	3	0.43
{B, D}	4	0.57
{C, D}	3	0.43

Los itemsets {A, B}, {B, C}, {B, D} y {C, D} superan el límite de soporte, por lo que son frecuentes. Los itemsets {A, C} y {A, D} no superan el soporte mínimo por lo que se descartan. Además, cualquier futuro itemset que los contenga también será descartado ya que no puede ser frecuente por el hecho de que contiene un subconjunto infrecuente.

El algoritmo Apriori

Itemset (k=3)
{A, B, C}
{A, B, D}
{B, C, D}
{C, D, A}

Se repite el proceso, esta vez creando *itemsets* de tamaño $k = 3$.
Los *itemsets* {A, B, C}, {A, B, D} y {C, D, A} contienen subconjuntos infrecuentes, por lo que son descartados. Para los restantes se calcula su soporte.

Itemset (k=3)	Ocurrencias	Soporte
{B, C, D}	2	0.29

El *items* {B, C, D} no supera el soporte mínimo por lo que se considera infrecuente. Al no haber ningún nuevo *itemset* frecuente, se detiene el algoritmo.

Itemset frecuentes
{A, B}
{B, C}
{B, D}
{C, D}

Como resultado de la búsqueda se han identificado los siguientes itemsets frecuentes

El algoritmo Apriori

El siguiente paso es crear las reglas de asociación a partir de cada uno de los *itemsets* frecuentes. De nuevo, se produce una explosión combinatoria de posibles reglas ya que, de cada *itemset* frecuente, se generan tantas reglas como posibles particiones binarias. En concreto, el proceso seguido es el siguiente. Supóngase que se desean únicamente reglas con una confianza igual o superior a 0.7, es decir, que la regla se cumpla un 70% de las veces. Finalmente se halla la confianza.

Reglas	Confianza	Confianza
{A} => {B}	soporte{A, B} / soporte {A}	0.43 / 0.43 = 1
{B} => {A}	soporte{A, B} / soporte {B}	0.43 / 0.86 = 0.5
{B} => {C}	soporte{B, C} / soporte {B}	0.43 / 0.86 = 0.5
{C} => {B}	soporte{B, C} / soporte {C}	0.43 / 0.57 = 0.75
{B} => {D}	soporte{B, D} / soporte {B}	0.43 / 0.86 = 0.5
{D} => {B}	soporte{B, D} / soporte {D}	0.43 / 0.71 = 0.6
{C} => {D}	soporte{C, D} / soporte {C}	0.43 / 0.57 = 0.75
{D} => {C}	soporte{C, D} / soporte {D}	0.43 / 0.71 = 0.6

De todas las posibles reglas, únicamente {A} => {B}, {C} => {D} y {C} => {B} superan el límite de confianza