

**t7-in.pdf**



**patrivc**



**Apuntes Variados**



**4º Grado en Ingeniería Informática**



**Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación  
Universidad de Granada**

**antes**



**Descarga sin publi  
con 1 coin**



**Después**

**WUOLAH**



Importante

Puedo eliminar la publi de este documento con 1 coin

¿Cómo consigo coins? → Plan Turbo: barato  
→ Planes pro: más coins

perdo  
espacio



Necesito  
concentración

ali ali ooh  
esto con 1 coin me  
lo quito yo...

WUOLAH

## Tema 7: Patrones frecuentes y reglas de asociación

### 1. Descubrimiento de asociaciones

Búsqueda de patrones frecuentes, asociaciones, correlaciones, o estructuras causales entre conjuntos de artículos u objetos (datos) a partir de bases de datos transaccionales, relacionales y otros conjuntos de datos. Búsqueda de secuencias o patrones temporales.

Aplicaciones:

- análisis de cestas de la compra (Market Basket analysis)
- diseño de catálogos,...
- ¿Qué hay en la cesta? Libros de Jazz
- ¿Qué podría haber en la cesta? El último CD de Jazz
- ¿Cómo motivar al cliente a comprar los artículos que es probable que le gusten?

#### 1.1. Market Basket Analysis (análisis de la cesta de la compra)

**Análisis de clientes:** se utiliza información sobre lo que ha comprado un cliente para ofrecernos una aproximación sobre quién es y por qué hace ciertas compras

**Análisis de productos:** aporta información sobre qué productos tienden a ser comprados juntos



Ejemplo: asociación de pañales y cervezas

Los clientes que compran cerveza también compran patatas  
¿Para eso no es necesario el uso de técnicas de Minería de Datos!  
Los viernes por la tarde, con frecuencia, quienes compran pañales, compran también cerveza.  
¿Qué significa? ¿A qué se debe?

### 2. Reglas de asociación

Se ha desarrollado una gran cantidad de investigación en torno al área de análisis de cestas de la compra, debido a la claridad y utilidad de sus resultados, que se expresan en forma de reglas de asociación.

Objetivo de los algoritmos de extracción de reglas de asociación:

- Dada una base de datos de transacciones, donde cada transacción es una lista de artículos (comprados por un cliente en la misma visita)
- Encontrar todas las reglas que co-relacionen la presencia de un conjunto de artículos con otro conjunto de artículos.
- Ejemplo: 98% de la gente que compra neumáticos y accesorios para el automóvil, también adquiere servicios (cambio de neumáticos, ...)

La idea es obtener reglas del tipo: “Antecedente => Consecuente [soporte, confianza]” en el ejemplo de pañales y cerveza es: compra(x, “pañales”) => compra(x, “cerveza”) [0.5%, 60%]

## 2.1 Conceptos básicos

Transacción:

Formato relacional

<Tid, item>

<1, item1>

<1, item2>

<2, item3>

Formato compacto

<Tid, itemset>

<1, {item1,item2}>

<2, {item3}>

- Item (o artículo): elemento individual
- Itemset (o conjunto): conjunto de items/artículos
- Soporte de un conjunto I: nº de transacciones conteniendo I
- Soporte mínimo  $m_s$ : umbral de soporte
- Conjunto frecuente: con soporte  $\geq m_s$

Los conjuntos frecuentes representan conjuntos de artículos que están correlacionados positivamente

## 2.2 Distintos tipos de reglas de asociación

- Asociaciones Booleanas vs Cuantitativas dependiendo del tipo de los valores que se manejan
  - compra (x, "SQLServer")  $\wedge$  compra (x, "Libro de MD")  $\rightarrow$  compra (x, "DBMiner") [0.2%, 60%]
  - Edad (x, '30..39')  $\wedge$  ingresos (x, '42K..48K')  $\rightarrow$  compra (x, 'PC') [1%, 75%]
- Asociaciones unidimensionales vs multidimensionales
  - A  $\rightarrow$  B A & B & ... & N  $\rightarrow$  D
- Análisis con distintos niveles de abstracción:
  - Edad (x, '30..39')  $\rightarrow$  compra (x, cerveza)
  - Edad (x, '30..39')  $\rightarrow$  compra (x, cerveza alemana)
- Posibles Extensiones:
  - Correlaciones, análisis de causalidad
- Asociación no implica necesariamente correlación o causalidad

## 2.3 Medidas de soporte y confianza

Encontrar todas las reglas  $X \& Y \Rightarrow Z$  con un mínimo de confianza y soporte





















- **Soporte (s)**: probabilidad de que una transacción contenga  $\{X \& Y \& Z\}$
- **Confianza (c)**: probabilidad condicional  $P(Z|X \& Y)$

Ejemplo: Sea el valor mínimo para confianza y soporte 50%:

transacción ID	artículos
1	A,B,C
2	A,C
3	A,D
4	B,E,F

$A \Rightarrow C$  (50%, 66.6%)

$C \Rightarrow A$  (50%, 100%)

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	  
Transaction 6	 
Transaction 7	 
Transaction 8	 

$$\text{Confidence} (\text{red circle} \rightarrow \text{yellow square}) = \frac{\text{Support} (\text{red circle, yellow square})}{\text{Support} (\text{red circle})}$$

**Coverage**: porcentaje de instancias que coinciden con el antecedente "A"

**Support**: porcentaje de instancias que coinciden con el antecedente "A" y el consecuente "C"








**Confidence**: porcentaje de instancias en el antecedente que también contienen el consecuente.

$$\text{Confidence} = \text{Support} / \text{Coverage}$$

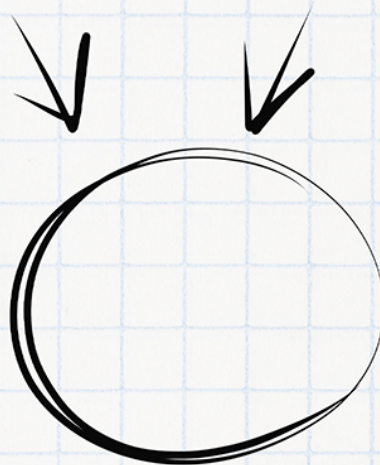


# Imagínate aprobando el examen

## Necesitas tiempo y concentración

Planes	 PLAN TURBO	 PLAN PRO	 PLAN PRO+
 Descargas sin publi al mes	10 	40 	80 
 Elimina el video entre descargas			
 Descarga carpetas			
 Descarga archivos grandes			
 Visualiza apuntes online sin publi			
 Elimina toda la publi web			
 Precios <span>Anual <input type="checkbox"/></span>	0,99 € / mes	3,99 € / mes	7,99 € / mes

Ahora que puedes conseguirlo,  
¿Qué nota vas a sacar?



# WUOLAH

## Apuntes Variados



Banco de apuntes de la

**WUOLAH**



**Comparte estos flyers en tu clase y consigue más dinero y recompensas**

- 1** Imprime esta hoja
- 2** Recorta por la mitad
- 3** Coloca en un lugar visible para que tus compis puedan escanar y acceder a apuntes
- 4** Llévate dinero por cada descarga de los documentos descargados a través de tu QR



## 2.4 Proceso de extracción

Al tratar con bases de datos grandes, el proceso se descompone en dos pasos:

1. Encontrar conjuntos de artículos frecuentes: Mayor ocurrencia que el soporte mínimo fijado
2. Generar reglas de asociación “fuerte” a partir de los conjuntos de artículos frecuentes: Deben satisfacer el mínimo fijado tanto para soporte como para confianza

Veamos un ejemplo:

transacción ID	artículos
1	A,B,C
2	A,C
3	A,D
4	B,E,F

Min. soporte: 50%
Min. confianza: 50%

frequent itemset	soporte
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

Transaction	Items
$t_1$	Bread,Jelly,PeanutButter
$t_2$	Bread,PeanutButter
$t_3$	Bread,Milk,PeanutButter
$t_4$	Beer,Bread
$t_5$	Beer,Milk

$X \Rightarrow Y$	s	$\alpha$
Bread $\Rightarrow$ PeanutButter	60%	75%
PeanutButter $\Rightarrow$ Bread	60%	100%
Beer $\Rightarrow$ Bread	20%	50%
PeanutButter $\Rightarrow$ Jelly	20%	33.3%
Jelly $\Rightarrow$ PeanutButter	20%	100%
Jelly $\Rightarrow$ Milk	0%	0%

Para la regla  $A \Rightarrow C$  :  
 soporte = soporte({ A & C }) = 50%  
 confianza = soporte({ A & C }) / soporte({ A }) = 66.6%

$I = \{\text{Beer, Bread, Jelly, Milk, PeanutButter}\}$   
 Soporte de {Bread,PeanutButter} es 60%

## 3. Algoritmo Apriori

Encuentra las asociaciones más frecuentes. Itera sobre la base de datos hasta que las asociaciones obtenidas no tienen el soporte mínimo. Método simple pero robusto. Salida intuitiva.

Los requisitos son:

- No necesita fijar los atributos de los lados derecho (consecuente) e izquierdo (antecedente) de las reglas pues se generan de manera automática
- Existen variedades para tratar todo tipo de datos
- Especificar mínimo soporte
- Especificar máximo número de reglas

El algoritmo busca iterativamente conjuntos frecuentes con cardinalidad 1 hasta k (k-conjunto), y después. Usa los conjuntos frecuentes para generar las reglas de asociación. En el paso clave del descubrimiento de **conjuntos frecuentes**, se basa en el principio “a priori”: cualquier subconjunto de un conjunto de artículos frecuente debe ser frecuente

Ejemplo: si {AB} es un conjunto frecuente, entonces tanto {A} como {B} deberían ser frecuentes

Esto permite definir el **principio de poda** en Apriori: Dado un conjunto “infrecuente”, no hay necesidad de generar sus superconjuntos (cualquier conjunto que contenga al subconjunto infrecuente, no es frecuente)

- Unión:  $C_k$  es generado uniendo conjuntos de  $L_{k-1}$  (se asume orden lexicográfico en las transacciones y que los prefijos son comunes)
- Poda: cualquier (k-1)-conjunto que no es frecuente, no puede ser un subconjunto de un k- conjunto frecuente

```

Ck : conjunto candidato de cardinalidad k
Lk : conjunto frecuente de cardinalidad k
L1 = {artículos frecuentes};
for(k=1;Lk !=∅;k++)do begin
    Ck+1 = candidatos generados desde Lk;
    for each transacción t en la base de datos do
        incrementar el contador de todos los candidatos en Ck+1 que están contenidos en t
    Lk+1 = candidatos en Ck+1 con min_support
end
return  $\bigcup_k L_k$ ;

```



Importante

Puedo eliminar la publi de este documento con 1 coin

¿Cómo consigo coins? → Plan Turbo: barato  
→ Planes pro: más coins

pierdo espacio



Generación de candidatos: suponemos que los ítems de  $L_k$  están ordenados:

### Paso 1: Unión sobre $L_k$

Insertar todo  $c$  en  $C_{k+1}$  tal que:  
 $c = \{p.item1, p.item2, \dots, p.itemk, q.itemk\}$   
donde  $p \in L_k$  y  $q \in L_k$  son tales que  
 $p.item1 = q.item1, \dots, p.itemk-1 = q.itemk-1$  y  
 $p.itemk < q.itemk$

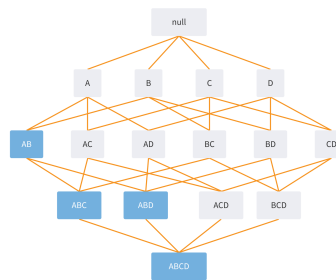
### Paso 2: Poda

Para todos los itemsets  $c$  de  $C_{k+1}$  hacer  
Para todos los  $k$ -subconjuntos  $s$  de  $c$  hacer  
Si ( $s \notin L_k$ ) entonces eliminar  $c$  de  $C_{k+1}$

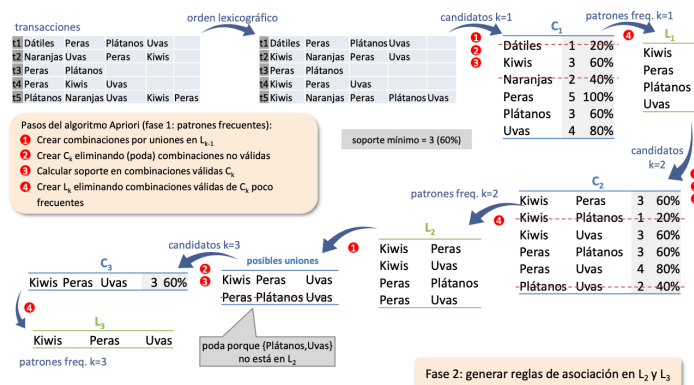
Ejemplo de generación de candidatos

- $L_3 = \{abc, abd, acd, ace, bcd\}$
- Unión en  $L_3: L_3 * L_3$ 
  - $abcd$  de  $ab-c$  y  $ab-d$
  - $acde$  de  $ac-d$  y  $ac-e$
- Poda:
  - $acde$  es eliminado porque  $ade$  y  $cde$  no están en  $L_3$
  - $abcd$  se conserva porque existen  $abc, abd, acd$  y  $bcd$
- $C_4 = \{abcd\}$

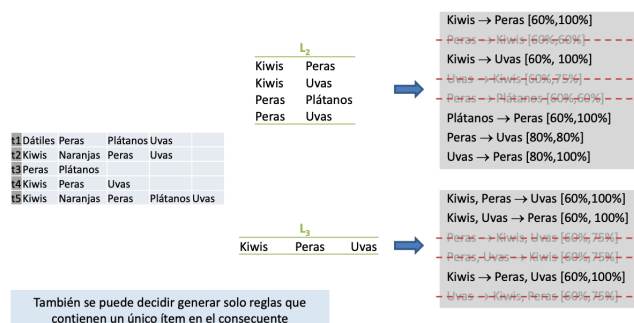
## 3.1 Espacio de búsqueda de conjuntos frecuentes



Si  $\{A,B\}$  no es frecuente, por el principio apriori, todos sus superconjuntos tampoco lo serán y se puede evitar analizarlos, reduciendo así el espacio de búsqueda



Fase 2: una vez que tenemos los conjuntos de elementos frecuentes  $L_k$ , podemos calcular la confianza y obtener las reglas que superan el umbral de confianza mínima

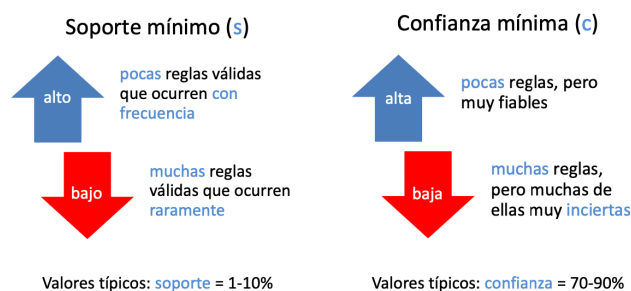


Necesito concentración

ali ali oohh esto con 1 coin me lo quito yo...

WUOLAH

### 3.2 Parámetros soporte y confianza



### 4. Medidas de interés

**Medidas objetivas:** soporte y confianza

**Medidas subjetivas:** Una regla (patrón) es interesante si es inesperada (sorprendente para el usuario) y/o útil (el usuario puede hacer algo con ella)

Veamos algunos ejemplos:

Críticas a Confianza

	basketball	not basketball	sum(row)
cereal	2000	1750	3750
not cereal	1000	250	1250
sum(col.)	3000	2000	5000

• Ejemplo 1:

- Entre 5000 estudiantes
  - 3000 juegan al baloncesto (60%)
  - 3750 comen cereales (75%), 1250 no comen cereales (25%)
  - 2000 juegan al baloncesto y comen cereales (40%)
- juega baloncesto*  $\Rightarrow$  *come cereales* [40%, 66.7%] no aporta nada porque el % global de alumnos que comen cereales es 75%, que es mayor que 66.7%
- juega baloncesto*  $\Rightarrow$  *no come cereales* [20%, 33.3%] sí aporta interés porque la confianza de la regla es 1.3333 (*lift*) veces mayor que solo la del consecuente

Críticas a Confianza

X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

Regla	Soporte	Confianza
$X \Rightarrow Y$	25%	50%
$X \Rightarrow Z$	37.50%	75%

• Ejemplo 2:

- X & Y: más correlacionadas
- X & Z: menos correlacionadas
- Sin embargo, confianza de  $X \Rightarrow Z$  domina
- El problema está en que la confianza se calcula sólo sobre el subconjunto de datos implicados en la regla, no se tiene en cuenta el total de datos. Se necesita una medida de dependencia o sucesos correlacionados
- $P(B|A)/P(B)$  se conoce como el **empuje** (*lift*) de la regla  $A \Rightarrow B$

**Interés** (correlación, empuje)

- $lift(A \rightarrow B) = P(B|A)/P(B) = P(A \text{ y } B)/(P(A) \cdot P(B))$
- Toma  $P(A)$  y  $P(B)$  en consideración
- $P(A \text{ y } B) = P(A) \cdot P(B)$  si A y B son independientes ( $lift = 1$ )
- A y B negativamente correlacionadas si *lift* es menor que 1; A y B positivamente correlacionadas si *lift* es mayor que 1

X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

Regla	Soporte	Lift
$X \rightarrow Y$	25%	2
$X \rightarrow Z$	37.50%	0.9
$Y \rightarrow Z$	12.50%	0.57

**Lift:** relación entre el apoyo del observador y el apoyo si A y C fueran estadísticamente independientes

$$\frac{Support}{p(A) * p(c)} = \frac{Confidence}{p(C)}$$

Si Lift < 1 Correlación negativa

Si Lift = 1 no hay correlacion

Si Lift > 1 correlacion positiva

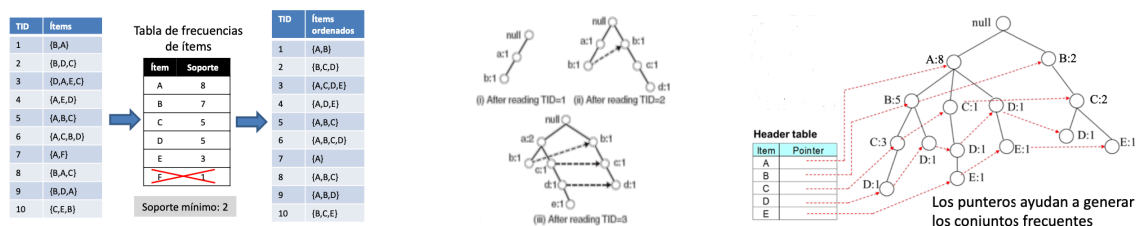


## 5. Otros algoritmos

### 5.1 FP-Growth

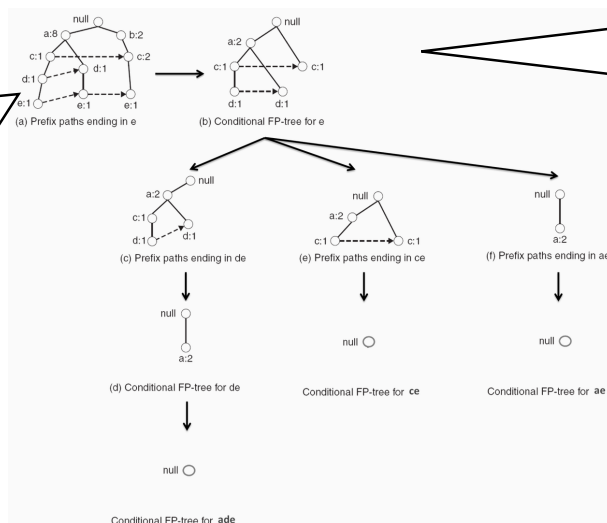
- FP-Growth (modelo basado en FP-tree): descubre frequent itemset sin generar los itemset candidatos
  - Fase 1.1: Construir una estructura compacta llamada FP-tree mediante dos pasadas al dataset
  - Fase 1.2: Extraer los frequent itemsets directamente del FP-tree:
- Encuentra todos los conjuntos de elementos frecuentes que terminan con un sufijo particular empleando una estrategia de divide y vencerás
  - Usando el puntero en la tabla de encabezado, descompone el árbol FP en varios subárboles, cada uno representa un subproblema según el sufijo
  - Para cada subproblema, recorre recursivamente el subárbol correspondiente de abajo hacia arriba para obtener bases de patrones condicionales

En la Fase 1.1 primero se calcula la tabla de frecuencias de ítems. Luego se construye el árbol siguiendo un orden en cada transacción de mayor a menor frecuencia de los ítems.



Fase 1.2. Ejemplo: base de patrones frecuentes para el sufijo 'e':

Soporte mínimo 2



Ramas que contienen 'e'

- Se ajustan las frecuencias a las de las hojas de cada rama (en este ejemplo, todo 1).
  - Se suman las ramas en los nodos que bifurcan varias ramas (por ejemplo,  $a:1+1=a:2$ ).
- Se eliminan los nodos que no alcanzan el soporte mínimo al sumar todas sus frecuencias, como pasa con 'b:1'. 'c' y 'd' se conservan por suman 2 entre sus dos nodos

Suffix	Frequent Itemsets
e	{e}, {d,e}, {a,d,e}, {c,e}, {a,e}
d	{d}, {c,d}, {b,c,d}, {a,c,d}, {b,d}, {a,b,d}, {a,d}
c	{c}, {b,c}, {a,b,c}, {a,c}
b	{b}, {a,b}
a	{a}

- OPUS search based algorithms
- Negative association rules
- Quantitative association rules (QAR)