Minería de datos

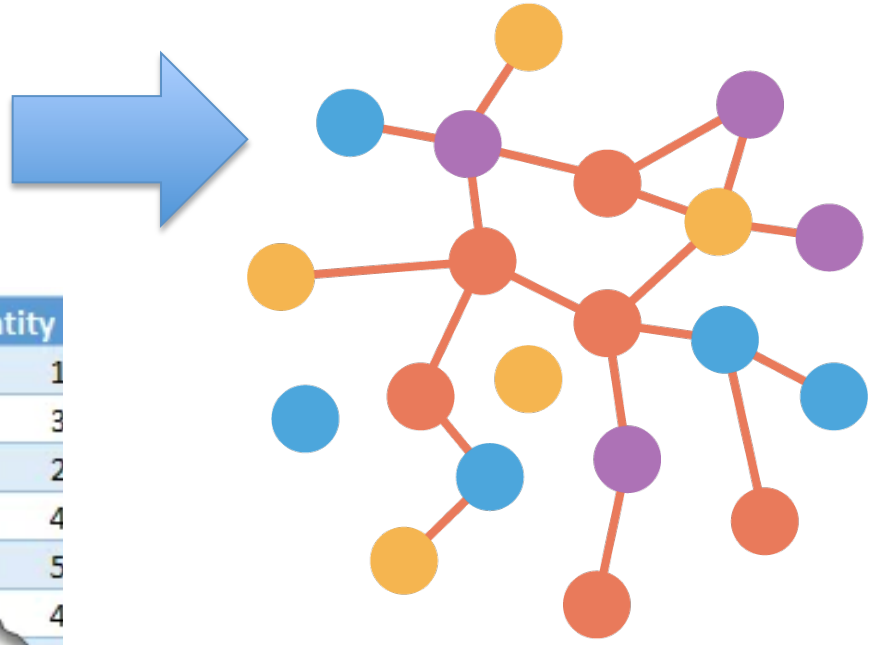
Reglas de asociación

Prof. Lorena Zúñiga S.

Febrero,2020

Reglas de asociación

Datos transaccionales



Sale ID	Time	Customer	Product ID	Quantity
S00001	12/1/2012 9:00:00 AM	C0001	P025	1
S00002	12/1/2012 9:05:58 AM	C0025	P025	3
S00003	12/1/2012 9:11:33 AM	C0010	P001	2
S00004	12/1/2012 9:17:16 AM	C0017	P023	4
S00005	12/1/2012 9:23:04 AM	C0018	P016	5
S00006	12/1/2012 9:28:43 AM	C0011	P018	4
S0007	12/1/2012 9:34:07 AM	C0015	P006	4

Análisis de asociación

- Es una rama del proceso de aprendizaje no supervisado
- Medir la fuerza de la co-ocurrencia entre dos items
- **No busca predecir** sino encontrar patrones

Análisis de asociación

Análisis de asociación - Usos

- análisis de transacciones (*retail*)
- análisis *clickstream*
- análisis de canasta de mercado
- sistemas recomendadores

¿Para qué le podría servir a un negocio?

Price bundling

Ventas cruzadas online

Análisis de asociación

- Considerar

Reglas de asociación

{Item A} -> {Item B}

Antecedente o
Premisa de la regla

Consecuencia o
Conclusión de la regla

- El antecedente y la consecuencia pueden contener más de 1 item

Minado de reglas de asociación

- Reglas de asociación básicas: ocurrencia de un item con otro
- Otras más complejas toman en cuenta:
  - cantidad de la ocurrencia
  - precio y secuencia de la ocurrencia, etc.

Pasos básicos

1. Preparar los datos → **formato de transacción**
2. Encontrar los conjuntos de items más frecuentes
3. Generar reglas de asociación relevantes a partir de los conjuntos de items

Formatos de transacción

Transacción	Items
1	{limones,pan, galletas}
2	{ayote, brócoli}
3	{jugo,galletas,leche}
4	{queso, jamón,vino,pan}

basket

Transacción	Item
1	{limones}
1	{pan}
1	{galletas}
2	{ayote}
2	{brócoli}
3	{jugo}
3	{galletas}
3	{leche}
4	{queso}
4	{jamón}
4	{vino}
4	{pan}

Minado de reglas de asociación

Ejemplo:

páginas visitadas en un sitio web

Minado de reglas de asociación

Datos iniciales

Session ID	Categorías accedidas
1	{Noticias, Finanzas}
2	{Noticias, Finanzas}
3	{Deportes, Finanzas, Noticias}
4	{Artes}
5	{Deportes, Noticias, Finanzas}
6	{Noticias, Artes, Entretenimiento}

Minado de reglas de asociación

- Los datos se transforman a un formato de transacción:

Session ID	Noticias	Finanzas	Entrete.	Deportes	Artes
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

Conjuntos de items

{Noticias, Finanzas} -> {Deportes}

¿Qué dice esta regla?

Conjuntos de items

- {Noticias, Finanzas} es un conjunto de items
- Puede existir en el antecedente o en la conclusión de la regla
- Deben ser conjuntos disjuntos (antecedente y conclusión)

Reglas de asociación

- **La fortaleza de una regla** de asociación se cuantifica mediante:
  - El soporte de un item, de la regla
  - La confianza
  - El lift

Soporte de un item

- Frecuencia relativa de ocurrencia de un conjunto de items en el conjunto de transacciones

Soporte de un item

Session ID	Noticias	Finanzas	Entrete.	Deportes	Artes
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

- **Ejemplo:**
  - Soporte({Noticias}) = 5/6 = 0.83
  - Soporte({Noticias, Finanzas}) = 4/6 = 0.67
  - Soporte({Artes}) = 2/6 = 0.33

Soporte de una regla

- cómo todos los items en una regla están representados en todas las transacciones
- indica si la regla es valiosa

Soporte de una regla

- **Ejemplo:**

  Regla: {Noticias} -> {Deportes},

  Los items {Noticias} y {Deportes} aparecen en 2 de 6 ocasiones, → el soporte de la regla es 0.33

Soporte de una regla

Session ID	Noticias	Finanzas	Entrete.	Deportes	Artes
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

Soporte de una regla

- ¿ Qué indica una regla con bajo soporte ?

- Se define un umbral, si se supera, la regla es considerada para análisis

Confianza de una regla

- Probabilidad de que ocurra la consecuencia en todas las transacciones que contienen el antecedente

Confianza(X -> Y) =

Soporte(X unión Y) / Soporte(X)

Confianza de una regla - Ejemplo

- Confianza({Noticias, Finanzas} -> {Deportes})

Session ID	Noticias	Finanzas	Entrete.	Deportes	Artes
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

Reglas de asociación

a) Soporte({Noticias, Finanzas, Deportes}) = 2/6
b) Soporte({Noticias, Finanzas}) = 4/6

Por lo tanto,

Confianza({Noticias, Finanzas} -> {Deportes}) =

- 2/6 **/** 4/6 = 0.5

¿Qué significa el valor 0.5 ?

Lift

- Incluye la frecuencia de ocurrencia de la conclusión de una regla

$$\textbf{Lift(X -> Y)} = \frac{\text{Soporte(X unión Y)}}{\text{Soporte(X) * Soporte(Y)}}$$

Lift - Ejemplo

- Lift({Noticias, Finanzas} -> {Deportes}) =

Soporte({Noticias, Finanzas, Deportes}) **/**
Soporte({Noticias, Finanzas}) * Soporte({Deportes})

2/6 / (4/6 * 2/6) =

0.333 / (0.666 * 0.333)

1.50

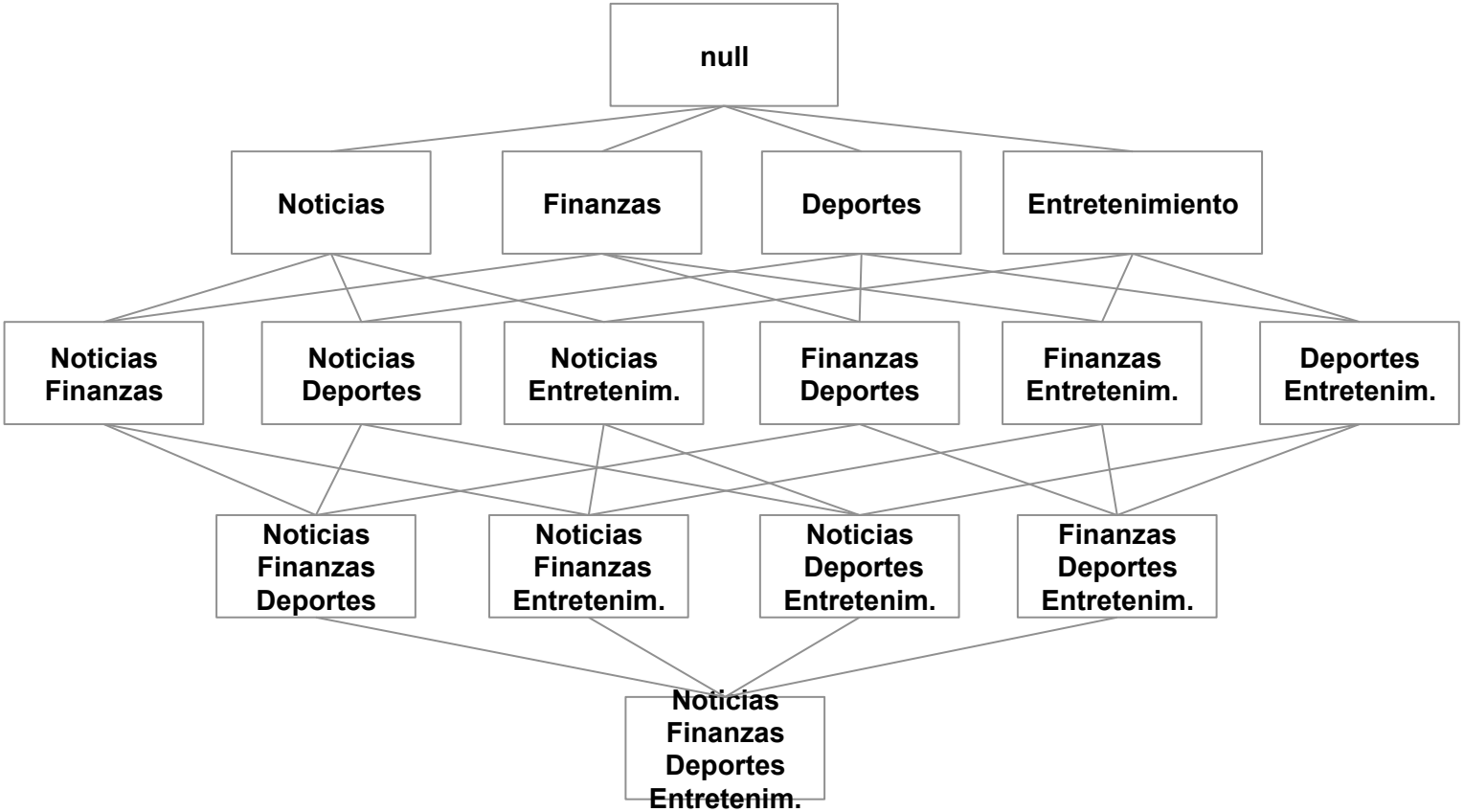
Lift

Valores cercanos a uno → que el antecedente y la consecuencia son independientes …

Generación de reglas

- **Proceso de generación de reglas**
  - Encontrar todos los conjuntos de items frecuentes ($2^n - 1$ conjuntos)
    - Árbol de conjuntos
  - Extraer reglas a partir de los conjuntos de items frecuentes ($3^n - 2^{n+1} + 1$)

Reglas de asociación



ALGORITMO A PRIORI

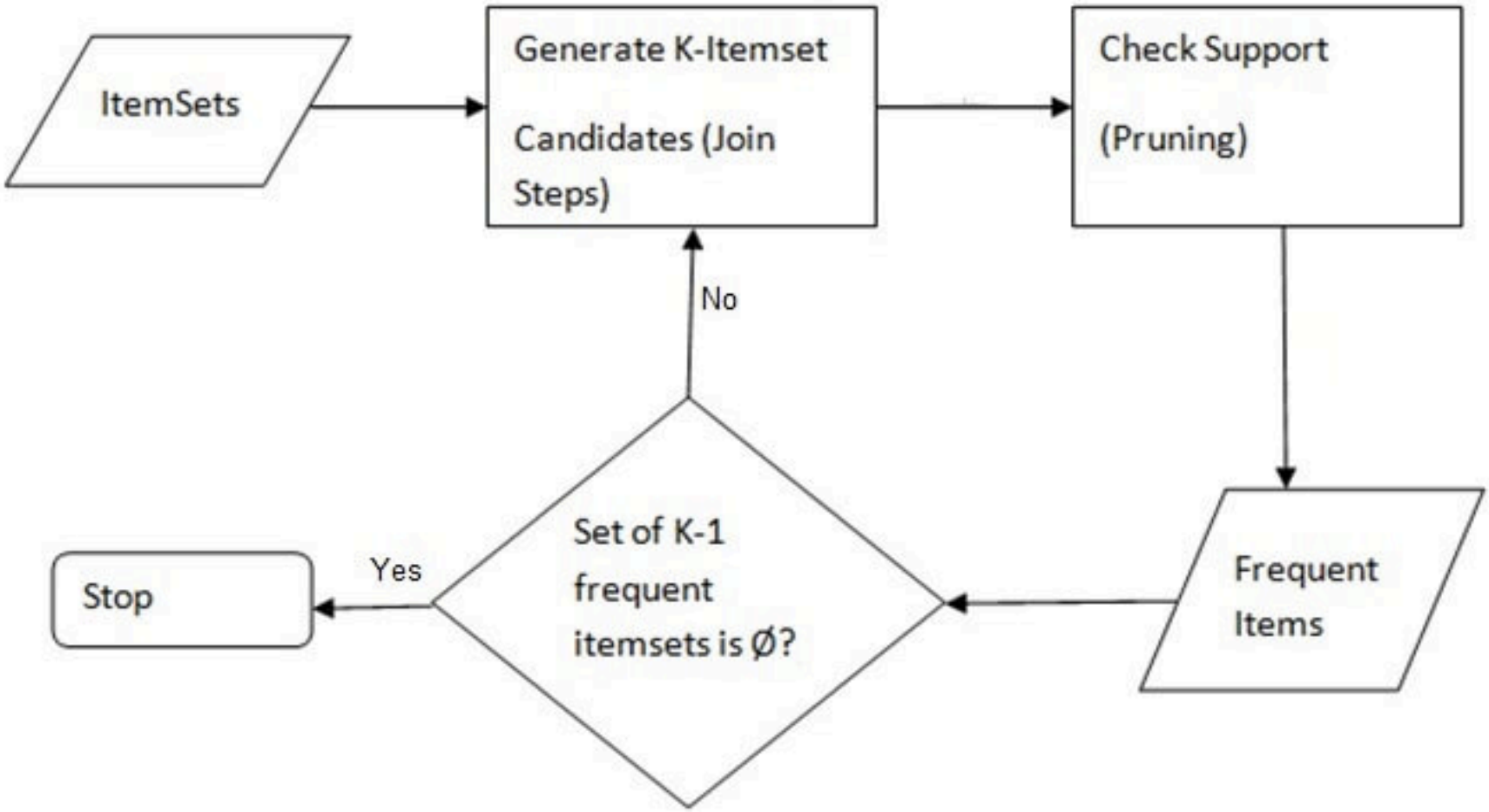
Propiedad A priori

Si un conjunto de elementos Z no es frecuente, entonces para cualquier elemento A, Z U A, será no frecuente

Algoritmo A priori

- Si un conjunto de elementos es frecuente, entonces todos sus subconjuntos lo serán
- ¿Cuándo es frecuente un elemento ?

Algoritmo A priori



Algoritmo A priori

- Se supone un umbral de soporte de 0.25
- Si {Noticias, Finanzas, Deportes} es frecuente, entonces todos sus subconjuntos lo serán:
  - {Noticias}
  - {Finanzas}
  - {Deportes}
  - {Noticias, Finanzas}
  - {Noticias, Deportes}

Algoritmo A priori

- Soporte({Noticias, Finanzas, Deportes}) = 0.33
  - Soporte({Noticias, Finanzas}) = 0.66
  - Soporte({Noticias, Deportes}) = 0.33
  - Soporte({Noticias}) = 0.83
  - Soporte({Finanzas}) = 0.66
  - Soporte({Deportes}) = 0.33

Algoritmo A priori

- Con los elementos infrecuentes sucede lo mismo …

Algoritmo A priori

- **Ejemplo:**
  - Soporte({Entretenimiento}) = 0.16
  - Revisando el soporte de sus subconjuntos:

Algoritmo A priori

- Soporte({Noticias, Entretenimiento}) = 0.16
- Soporte({Finanzas, Entretenimiento}) = 0
- Soporte({Deportes, Entretenimiento}) = 0
- Soporte({Noticias, Finanzas, Entretenimiento}) = 0
- Soporte({Noticias, Deportes, Entretenimiento}) = 0
- Soporte({Finanzas, Deportes, Entretenimiento}) = 0
- Soporte({Noticias, Finanzas, Deportes, Entretenimiento}) = 0

Algoritmo A priori

- A partir de estas visitas genere los conjuntos frecuentes.
- Suponga un umbral de soporte de 0.25

Session ID	Noticias	Finanzas	Entretenim.	Deportes
1	1	1	0	0
2	1	1	0	0
3	1	1	0	1
4	0	0	0	0
5	1	1	0	1
6	1	0	1	0

Algoritmo A priori

- **¿Cuántos posibles conjuntos hay?**
  - $R/2^n - 1$ posibles conjuntos = $2^4 - 1 = 15$
  - Soporte de conjuntos de 1 elemento:

Elemento	Conteo	Soporte
{Noticias}	5	0.83
{Finanzas}	4	0.66
{Entretenimiento}	1	0.17
{Deportes}	2	0.33

Algoritmo A priori

- Como Soporte({Entretenimiento}) < 0.25 , se descarta, sus subconjuntos también
- Se repite el proceso para conjuntos de 2 elementos

Elemento	Conteo	Soporte
{Noticias, Finanzas}	4	0.66
{Noticias, Deportes}	2	0.33
{Finanzas, Deportes}	2	0.33

Algoritmo A priori

- Se calcula el conteo de soporte y el soporte de cada transacción o visita para conjuntos de 3 elementos:

Elemento	Conteo	Soporte
{Noticias, Finanzas, Deportes}	2	0.33

- El proceso continua hasta generar todos los conjuntos de n elementos

Algoritmo A priori

- **Generación de reglas**
- Utilidad aproximada por una medida (confianza, lift, o convicción)
- 2^n – 2 reglas por conjunto frecuente

Algoritmo A priori

- **Ejemplo**
  - Del conjunto {Noticias, Finanzas, Deportes} se podrían generar las siguientes reglas:
    - {Noticias, Deportes} -> {Finanzas} 0.33/0.33 = **1**
    - {Noticias, Finanzas} -> {Deportes} 0.33/0.66 = **0.5**
    - {Deportes, Finanzas} -> {Noticias} 0.33/0.33 = **1**
    - {Noticias} -> {Deportes, Finanzas} 0.33/0.83 = **0.4**
    - {Deportes} -> {Noticias, Finanzas} 0.33/0.33 = **1**
    - {Finanzas} -> {Noticias, Deportes} 0.33/0.66 = **0.5**

Algoritmo A priori

- **Generación de reglas**
- Todas las reglas que sobrepasen un límite de confianza se toman para análisis

Ejemplo en R

ALGORITMO FP-GROWTH

FP-Growth

- ***Frequent Pattern Growth***
- Calcula los conjuntos de elementos frecuentes usando un grafo, llamado FP-Tree
- Usa un árbol comprimido para generar los conjuntos de elementos frecuentes

FP-Growth

- Transformar los datos en un grafo, representando las rutas frecuentes
- Se ordenan descendentemente las transacciones por frecuencia

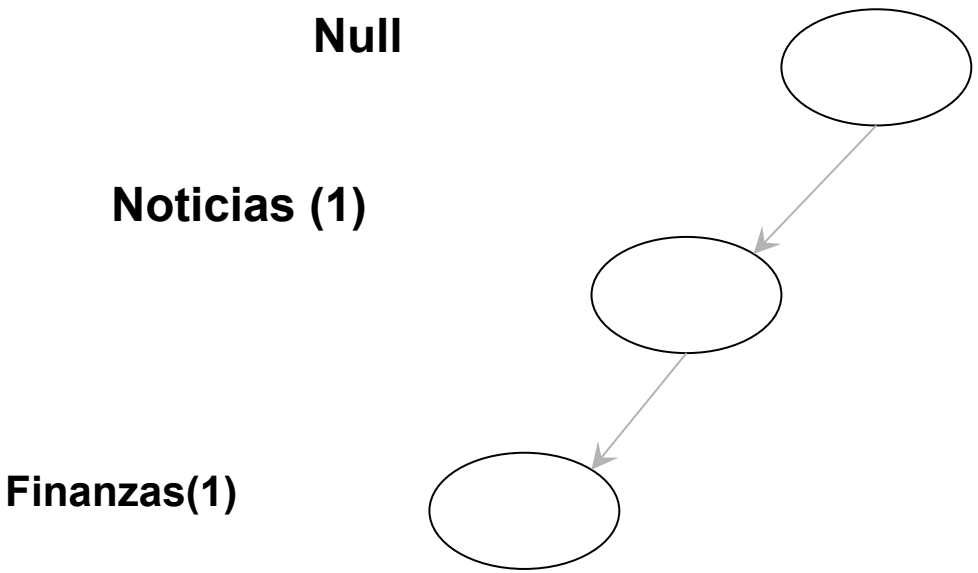
FP-Growth

- Ejemplo

Session ID	Categorías accedidas
1	{Noticias, Finanzas}
2	{Noticias, Finanzas}
3	{Deportes, Finanzas, Noticias}
4	{Deportes}
5	{Deportes, Noticias, Finanzas}
6	{Noticias, Entretenimiento}

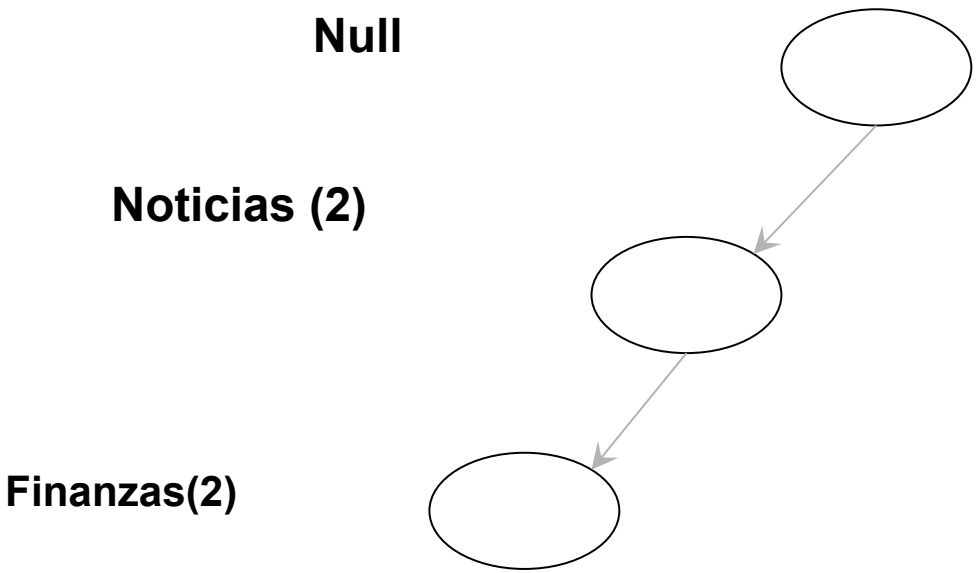
FP-Growth

- Mapeo de la 1ª transacción: {Noticias, Finanzas}



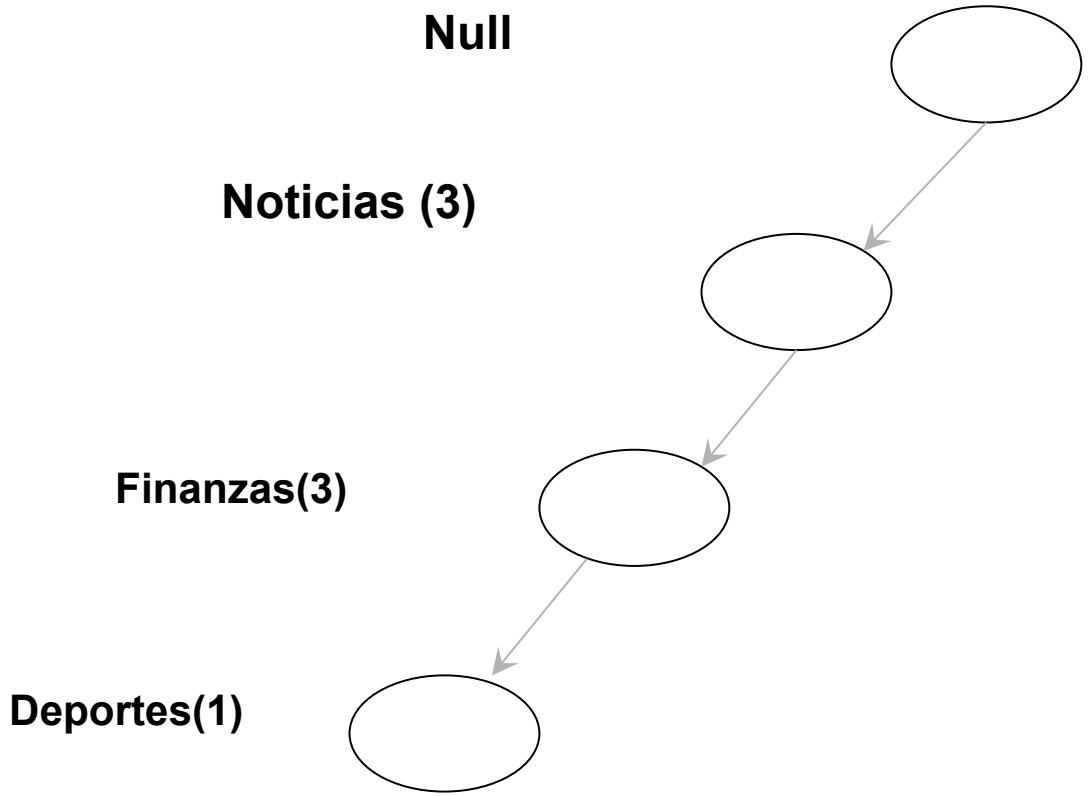
FP-Growth

- Mapeo de la 2ª transacción: {Noticias, Finanzas}



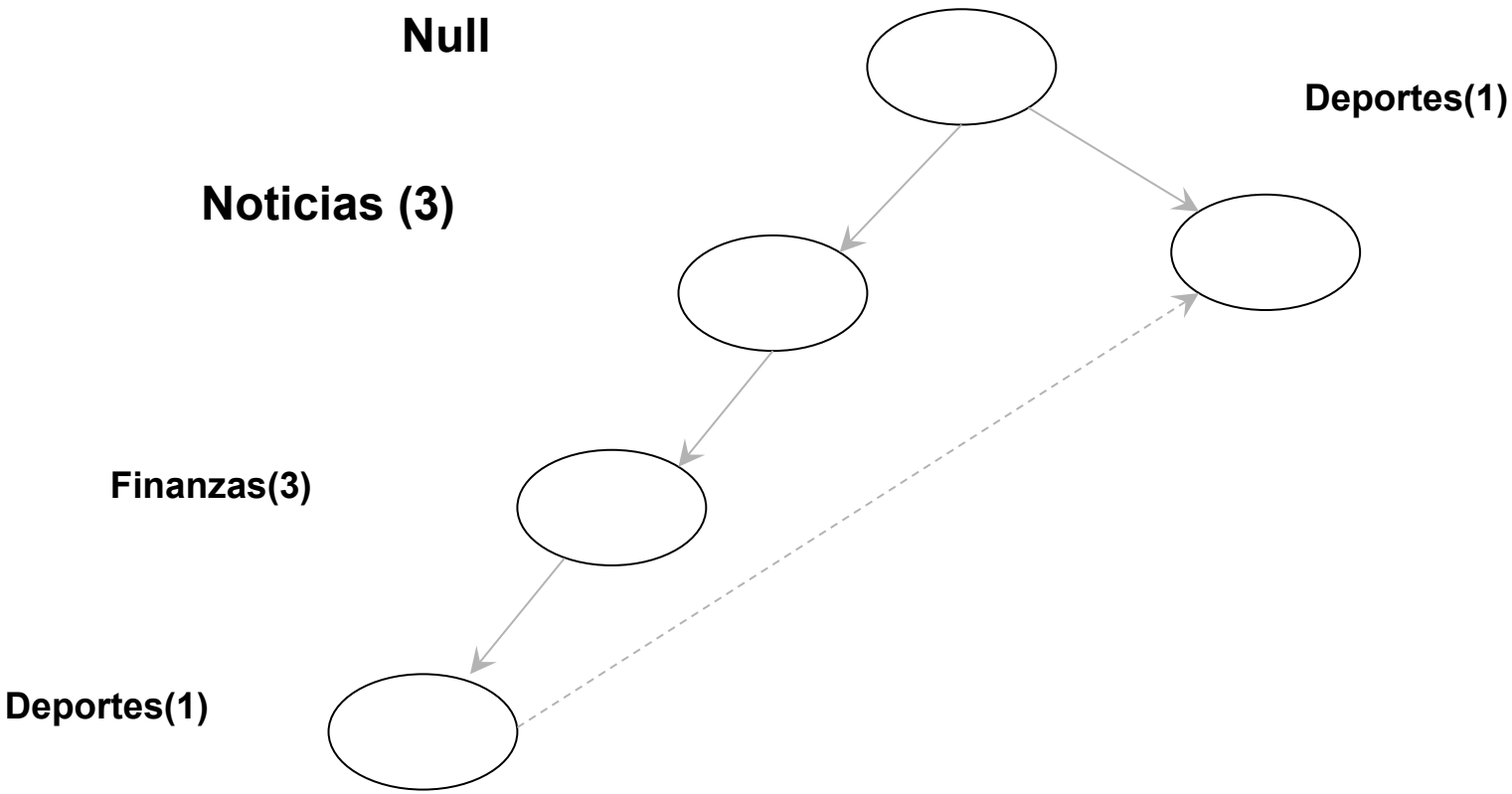
FP-Growth

- Mapeo de la 3ª transacción: {Noticias, Finanzas, Deportes}



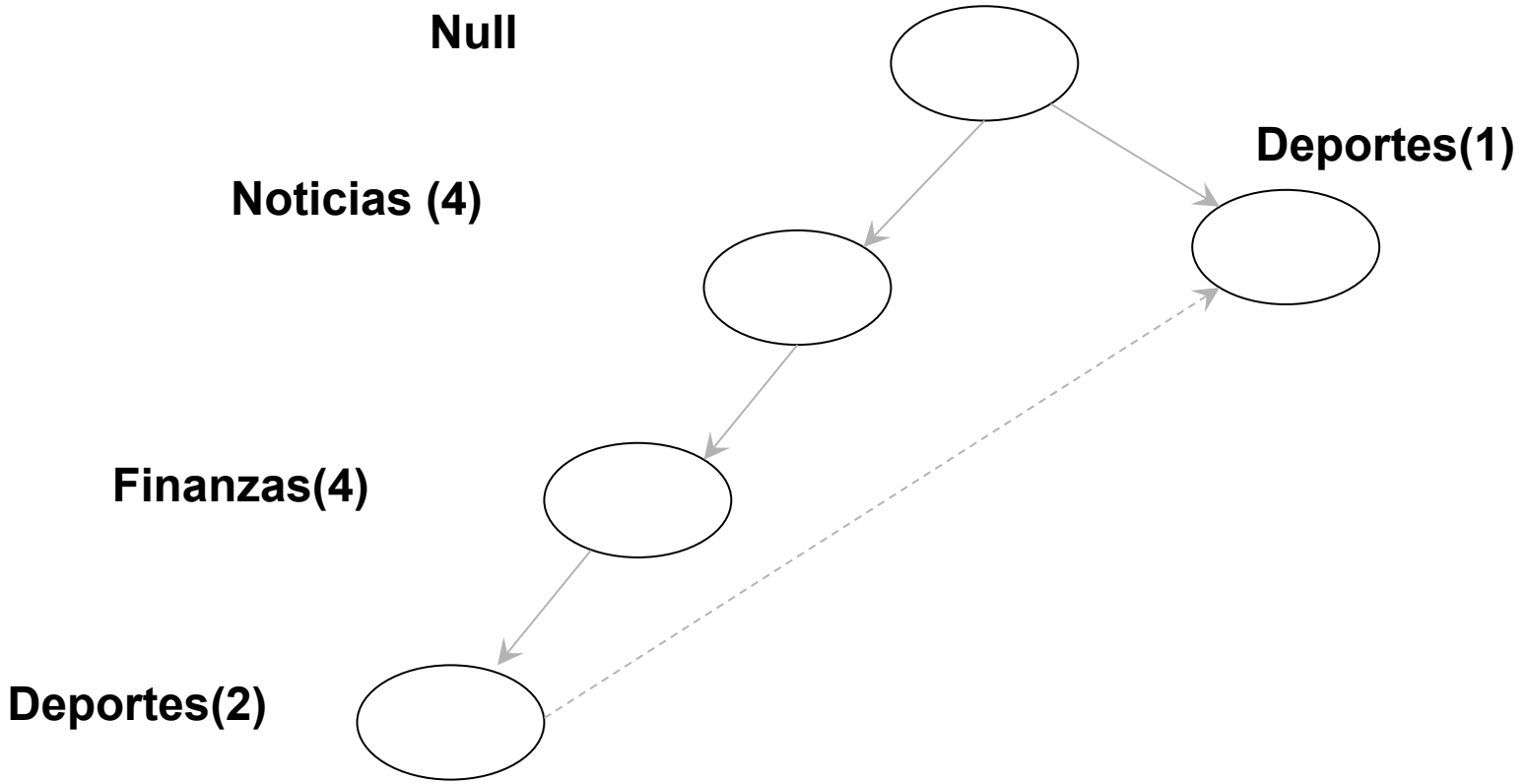
FP-Growth

- Mapeo de la 4ª transacción: {Deportes}



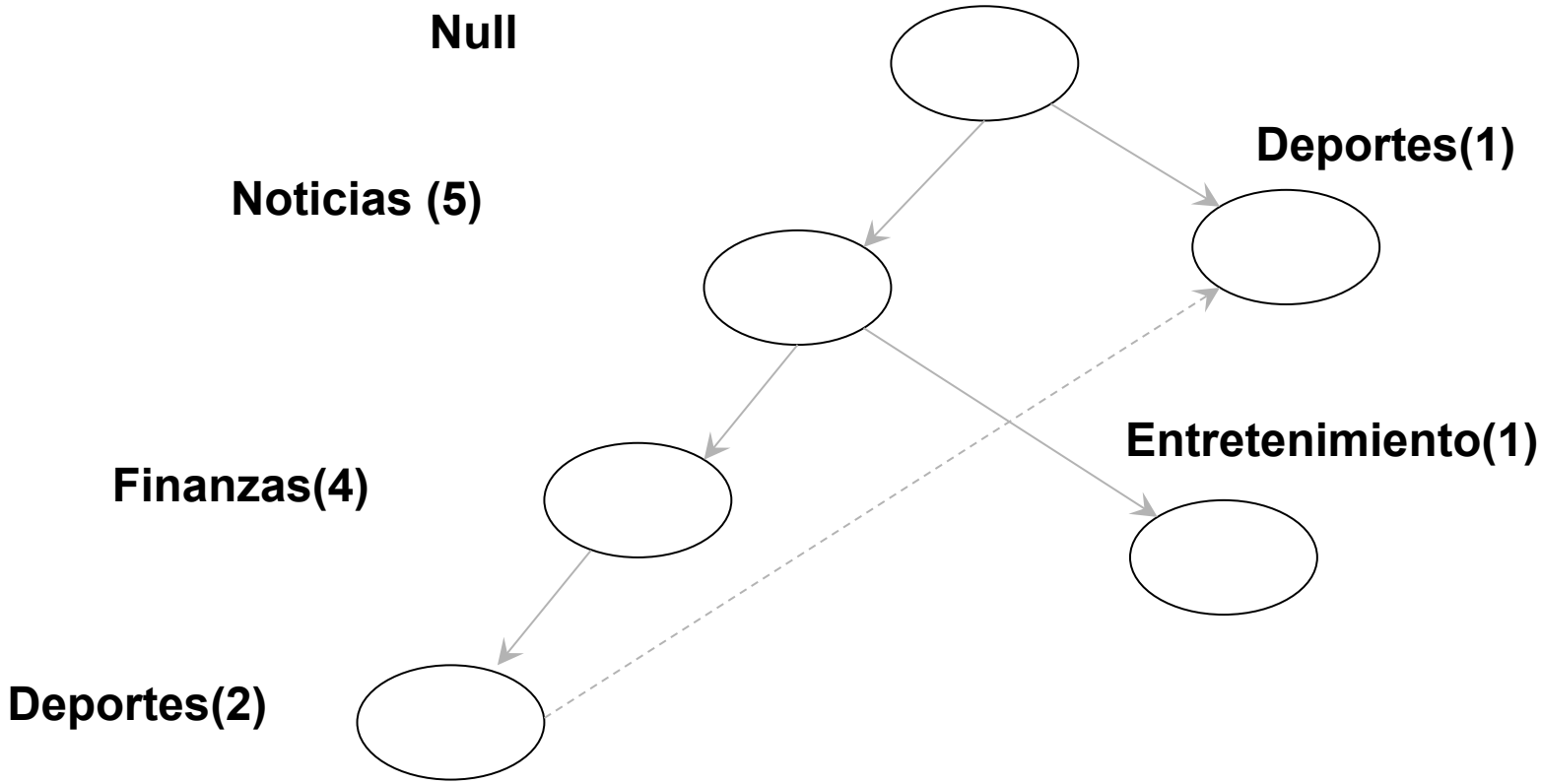
FP-Growth

- Mapeo de la 5ª transacción: {Noticias, Finanzas, Deportes}



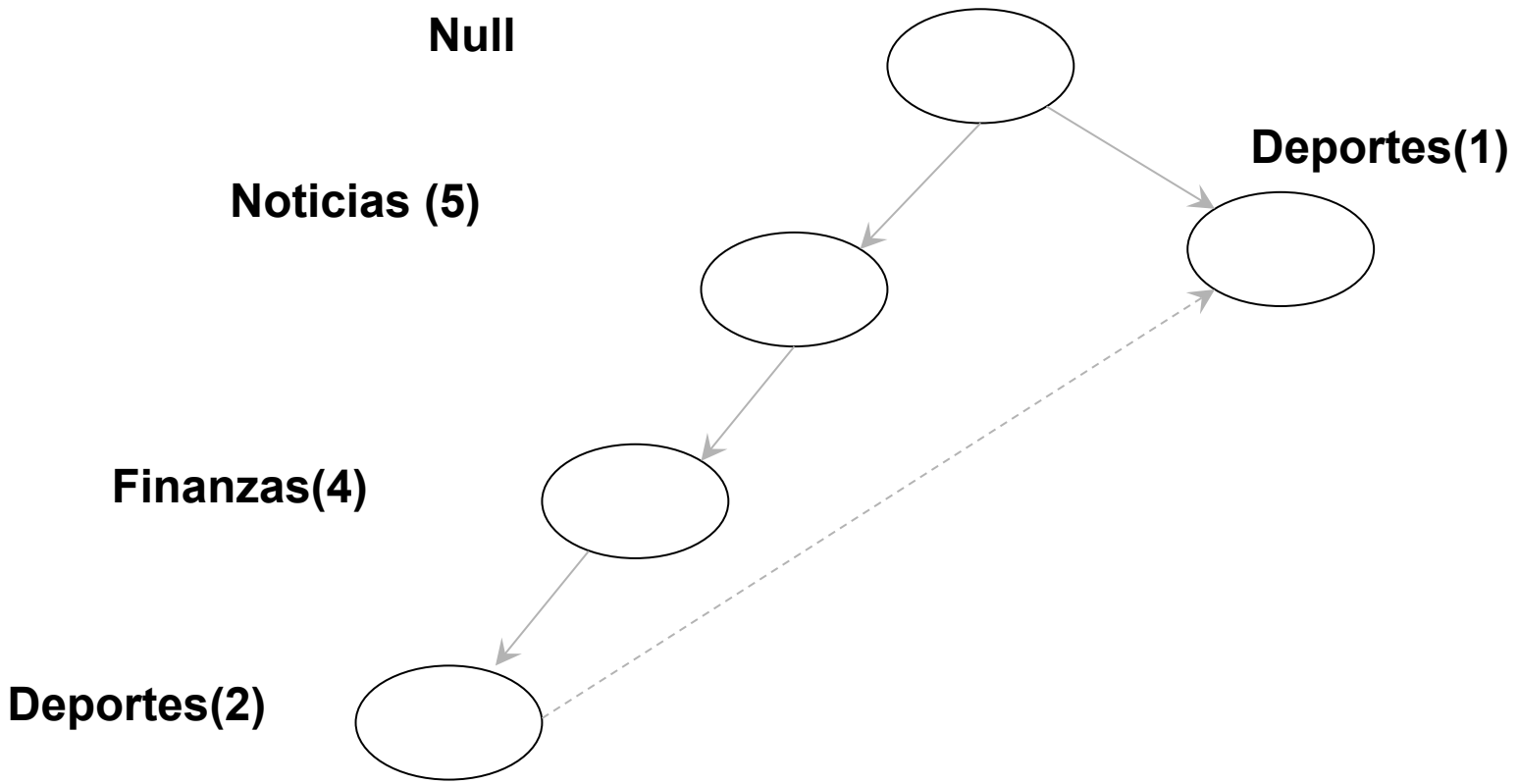
FP-Growth

- Mapeo de la 6ª transacción: {Noticias, Entretenimiento}



FP-Growth

- Árbol/grafo comprimido



FP-Growth

- Empieza con los elementos menos frecuentes:
  - Si el soporte del elemento > límite establecido, busca todos que terminen con ese elemento.
  - Ejemplo:
    - buscar todos los que finalicen con {Deportes}

Ejemplo en RapidMiner

¿Preguntas?