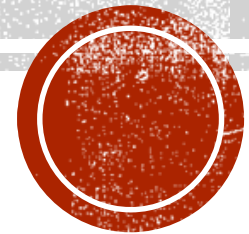


APRENDIZAJE NO SUPERVISADO



Javier Diaz Cely, PhD

AGENDA



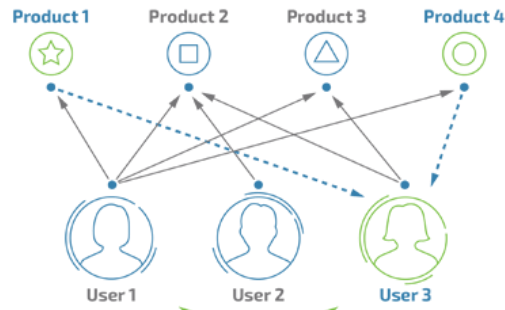
Aprendizaje automático



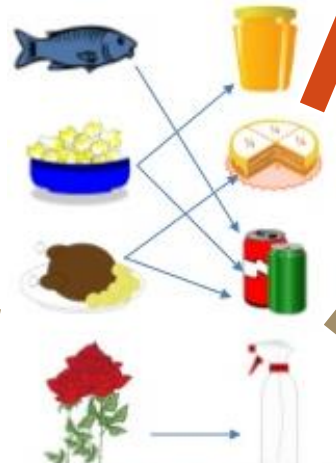
Aprendizaje supervisado



Aprendizaje no supervisado



Filtro colaborativo



Sistemas de recomendación

$$tfidf_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

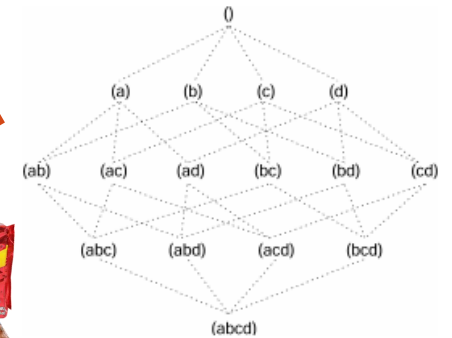
TF/IDF



Basados en contenido



Itemset



Apriori



SISTEMAS DE RECOMENDACIÓN



<http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>

SISTEMAS DE RECOMENDACIÓN

Cervezas y pañales

- Mito o leyenda?
- 1992 USA cadena de supermercados. Wal-mart?
No, OSCO
- Hombres jóvenes que compran **pañales** entre las 5 y las 6 de la tarde los viernes y sábados, también compran **cerveza**

→ Posicionar la cerveza cerca de los pañales implicó un crecimiento en su venta



SISTEMAS DE RECOMENDACIÓN

TARGET sabe que estás embarazada

- Target Megastore
- Hábitos de compra arraigados
- Grandes cambios presentan una ventana de vulnerabilidad
- Hay que llegar primero



SISTEMAS DE RECOMENDACIÓN

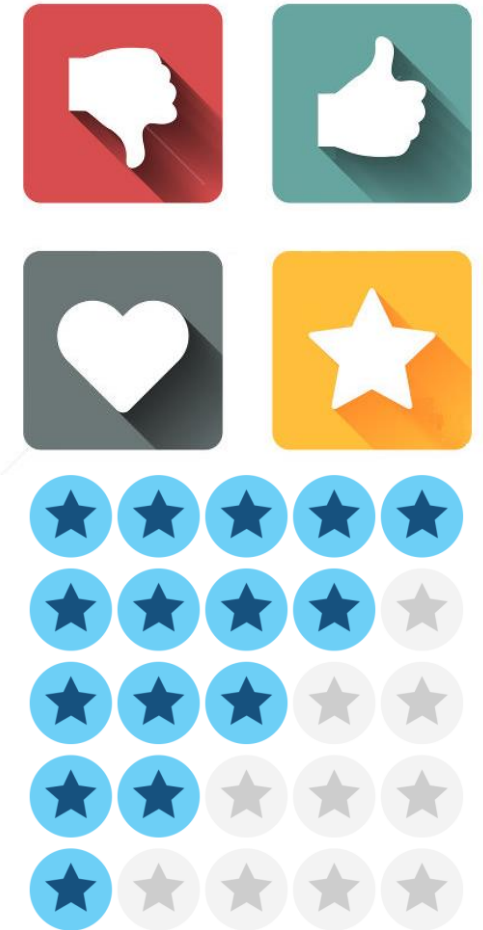
NETFLIX Prize, 2006 a 2011

- 1 millón de dólares
- Mejorar su sistema de recomendación (Cinematch), diseñado para predecir si una película le gustaría a un usuario, dados sus calificaciones de otras películas
- Proveen datos anónimos (no lo fueron tanto) con los ratings de las películas
- Objetivo: mejorar en 10% la predicción (accuracy) de Cinematch
- <http://www.netflixprize.com/rules.html>



SISTEMAS DE RECOMENDACIÓN

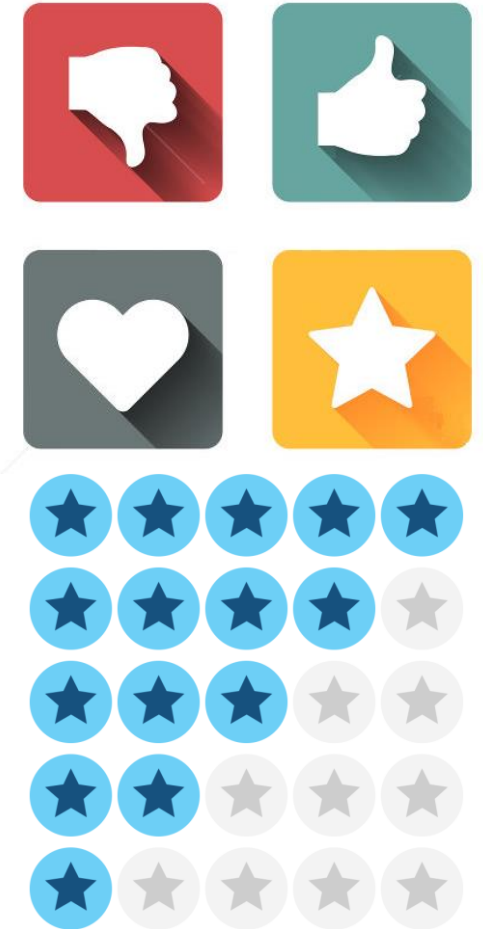
- **Un cliente es un activo.** Se puede predecir su comportamiento de compra:
 - Preferencias, intereses, interacciones, socio-demográficos
 - Identificación: tarjetas de fidelidad, cédula, email, IP
 - Para maximizar ventas (perspectiva de ventas al detal)
 - Para mejorar la satisfacción propia (perspectiva del cliente)
- **Datos: matriz de transacciones de clientes por ítems**
 - Cada fila es un cliente, cada columna es un ítem (producto)
 - Almacenamiento de información unaria, binaria, cantidades, ratings
 - Matriz puede ser muy grande
 - Matriz puede ser muy dispersa
 - Captura explícita o implícita



SISTEMAS DE RECOMENDACIÓN

Datos utilizados para la recomendación

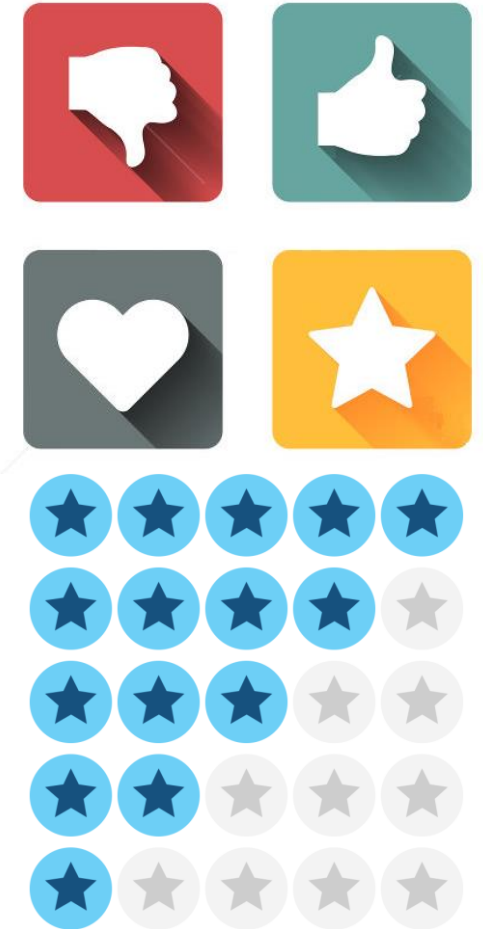
- **Explícitos:** preguntar a los clientes qué opinan de cada producto. Diferentes tipos:
 - Ratings:
 - Escalas (1-5, -10 to +10)?
 - Discretos o continuos (e.g. media estrella)?
 - Escala unaria: ítems con corta vida útil (e.g. noticias)
 - Escala binaria: votos +/- . Controversia es posible
 - Escala binaria a partir de umbral: (e.g. ≥ 4 estrellas)



SISTEMAS DE RECOMENDACIÓN


Datos utilizados para la recomendación

- **Implícitos:** inferidos a partir de las acciones de los clientes, de diferentes tipos:
 - Compras, accesos, lecturas, clicks, tiempo dedicado a un ítem
 - Análisis de sentimiento de los reviews
 - Omisión de acción (e.g. click en el 3er ítem presentado)
- Preferencias pueden cambiar con el tiempo
- Diferentes significados de las escalas para diferentes personas

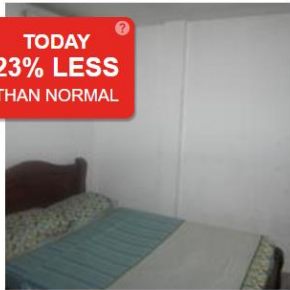


NO PERSONALIZADOS


- Nada acerca del consumidor
- Agregación de opiniones
 - Promedio, mediana → pérdida de información
 - Distribución porcentual por rating
 - Rankings objetivos (e.g. taquilla)
- Problemas:
 - Información vieja
 - Sesgo de auto selección
 - Diversidad de opinión



Hotel Maroel ★★
Getsemani, Cartagena (0.3 km from center)
Last booked: 8 hours ago
Value Deal
Double Room
In high demand – only 2 rooms left!
COP 45,000
Review score 5.0
151 reviews
Select your room >



Hotel Dora Smith Cartagena
Cartagena (5 km from center)
2 people are looking right now
Reservation possible without a credit card
Booked 10 times today
Twin Room
We have 3 rooms left!
~~-23% COP 60,000~~ **COP 46,000**
Good 7.7
11 reviews
Select your room >



Hotel LM A Luxury Boutique Hotel ★★★★★
Centro, Cartagena (0.5 km from center)
1 person is looking right now
Booked 8 times in the last 48 hours
You missed it! We reserved our last available room at this property.
Exceptional 9.5
Location 9.9
205 reviews



NO PERSONALIZADOS

- Basados en popularidad:
 - ¿Qué ítems se están consumiendo en este momento?
 - Los que compraron este ítem también compraron...
 - Considerando tus compras actuales, deberías comprar esto...
- Reglas de asociación:
 - Probabilidad condicional de comprar un ítem dados los productos ya comprados
 - Problema de reglas obvias (pan, arroz)
 - Problema de productos poco comprados (anchoas)
 - Posible no consideración de productos no comprados

People who liked this also liked... [Learn more](#)

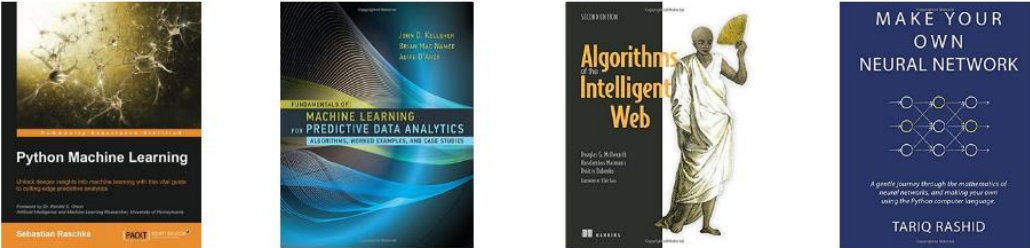


The Godfather (1972)
R Crime | Drama
★★★★★ 9.2/10
The aging patriarch of an organized crime dynasty transfers control of his clandestine empire to his reluctant son.
Director: Francis Ford Coppola
Stars: Marlon Brando, Al Pacino, Ja...

Add to Watchlist
Next »

◀ Prev 6 Next 6 ▶

Customers Who Bought This Item Also Bought



Python Machine Learning
Sebastian Raschka
★★★★★ 74

Fundamentals of Machine Learning for Predictive Data Analytics
James D. Kelleher, Brian H. Houghton, Niall Murtagh

Algorithms of the Intelligent Web
Douglas McIlwraith

Make Your Own Neural Network
Tariq Rashid



SISTEMAS DE RECOMENDACIÓN

- Sistemas no personalizados
 - Reglas de asociación
- Basados en contenido
 - TF/IDF
- Filtro colaborativo
 - Basado en usuarios
 - Basado en ítems



AGENDA



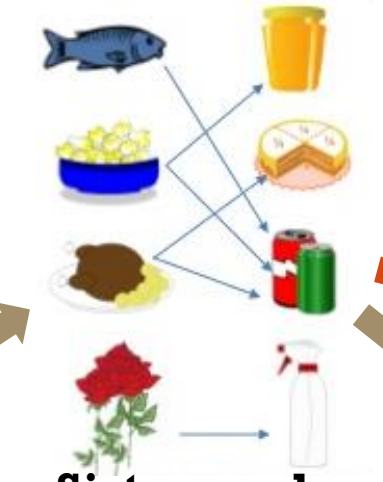
**Aprendizaje
automático**



**Aprendizaje
supervisado**



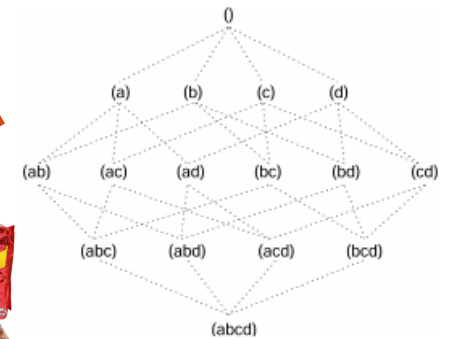
**Aprendizaje
no supervisado**



**Sistemas de
recomendación**



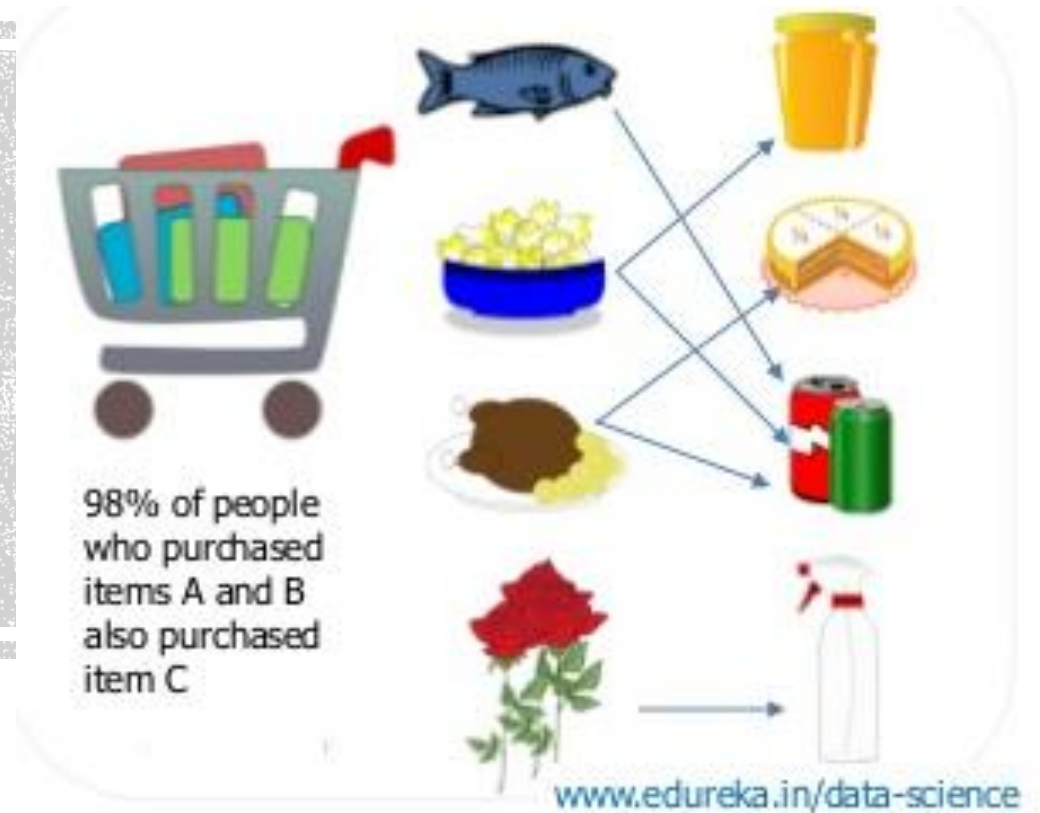
Itemset



Apriori



REGLAS DE ASOCIACIÓN — APRIORI



REGLAS DE ASOCIACIÓN

- **Market basket analysis (Análisis del carro de mercado):**
 - ¿Cuáles son los ítems mas propensos a ser comprados en conjunto? ...en los próximos 3 meses?
 - ¿Qué compran las personas con gustos similares?
 - Ofertas, amarres de productos, posición en el estante, venta cruzada
- **Predicción de navegación Web:**
 - Análisis de clickstream: ¿Cuál es el siguiente ítem más propenso a ser clickeado, o página a ser visitada?
- **Multimedia:**
 - Identificación de objetos en imágenes, videos o media social
 - Encontrar frases, entidades o atributos importantes en textos de gran volumen
- **Bioteología**
 - Encontrar secuencias de proteínas repetidas en secuencias genómicas del DNA
- **Social Networks**
 - Encontrar comunidades escondidas



REGLAS DE ASOCIACIÓN

- Aprendizaje no supervisado para descubrir **relaciones** significativas escondidas en el dataset
- **Transacción:** lista de productos comprados en conjunto en una misma visita a la tienda
- **Itemset:** Conjunto de uno o más productos
- **Itemset frecuente:** itemset cuyos ítems son frecuentemente comprados juntos (con respecto a un nivel mínimo de **soporte**)
- **Soporte:** Fracción de las transacciones que contienen un itemset dado → absoluta (conteo) o relativa (porcentaje)
- **Reglas:** itemset A → itemset B
- **Conocimiento de dominio:** Algunas reglas descubiertas pueden resultar inútiles por su obviedad (Papel → Lápiz), otras pueden resultar inesperadas, por tanto útiles (Pañal → Cerveza)

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

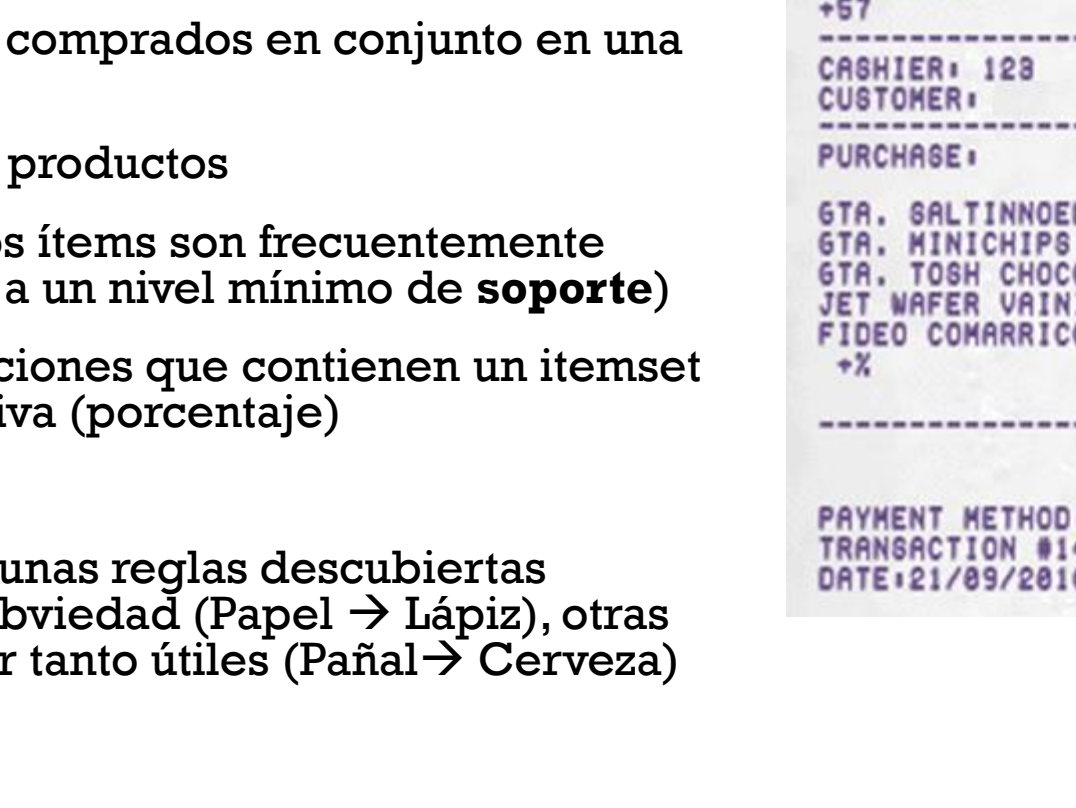
CASHIER: 123
CUSTOMER: **JUAN PEREZ**

PURCHASE:

GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
JET WAFER VAINILLA 20PLEX	\$2.00
FIDEO COMARRICO CLASICA X	\$4.00
+%	
TAX:	\$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024200 -001
DATE: 21/09/2016 5:01:29 PM



REGLAS DE ASOCIACIÓN — MEDIDAS

- **Soporte** $(A \rightarrow B) = P(A \ \& \ B)$, **simétrica**
- **Confianza** $(A \rightarrow B) = P(B \mid A) = P(A \ \& \ B) / P(A)$, **asimétrica**
 - Probabilidad condicional
 - Indica qué tanto se puede confiar en la regla, pero no si se trata de una coincidencia
- **Lift** $(A \rightarrow B) = \text{Confianza}(A \rightarrow B) / P(B) = P(A \ \& \ B) / P(A) * P(B)$, **simétrica**
 - Cuántas veces más ocurren A y B juntas que lo que se esperaría si fueran independientes
 - $=1$: Regla inútil. A y B son **independientes** entre ellas (no hay relación significativa)
 - >1 : La regla es útil. Entre mayor el lift mejor. Se trata de productos **complementarios**
 - <1 : La regla es útil para identificar productos **sustitutos**
- **Leverage** $(A \rightarrow B) = P(A \ \& \ B) - P(A) * P(B)$, **simétrica**
 - Medida análoga al lift, pero aditiva, y utilizando 0 como el límite de decisión



REGLAS DE ASOCIACIÓN — ITEMSET FRECUENTE



1-Itemset



4-Itemset



REGLAS DE ASOCIACIÓN — ITEMSET FRECUENTE

¿Cuál es el soporte del itemset {Café Sello Rojo}?

DISFRUTAR LA VIDA TE ALIMENTA
GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLIN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **XXX**

PURCHASE:

GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
JET WAFER VAINILLA 20PLEX	\$2.00
FIDEO COMARRICO CLASICA X	\$4.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024200 -001
DATE:21/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA
GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLIN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **YYY**

PURCHASE:

CAFE SELLO ROJO MEDIO 250	\$3.00
SPAGHETTI DORIA CLASICA X	\$3.00
CHOCOL. DIANA 500Gx16PAST	\$2.00
FIDEO COMARRICO CLASICA X	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
+%	TAX: \$0.00

TOTAL: \$15.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024266 -001
DATE:11/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA
GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLIN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **ZZZ**

PURCHASE:

GTA. TOSH MIEL BS. 9X3	\$3.00
CAFE SELLO ROJO MEDIO 250	\$2.00
SPAGHETTI DORIA CLASICA X	\$4.00
GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
+%	TAX: \$0.00

TOTAL: \$15.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475023934 -001
DATE:27/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA
GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLIN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **PPP**

PURCHASE:

CAFE SELLO ROJO MEDIO 250	\$4.00
PASAB. LA ESPECIAL SAL BP	\$2.00
GTA. DUCALES TC. X6 720G	\$3.00
CHOCOL.CHOCOLYNE CLAS 6PL	\$5.00
JET WAFER VAINILLA 20PLEX	\$2.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024073 -001
DATE:24/09/2016 5:01:29 PM

Soporte de {Café Sello Rojo} = $3/4 = 75\%$



REGLAS DE ASOCIACIÓN — ITEMSET FRECUENTE

¿Cuál es el soporte del itemset {Spaghetti Doria Clásica}?

DISFRUTAR LA VIDA TE ALIMENTA
GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **XXX**

PURCHASE:

GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
JET WAFER VAINILLA 20PLEX	\$2.00
FIDEO COMARRICO CLASICA X	\$4.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024200 -001
DATE:21/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA
GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **YYY**

PURCHASE:

CAFE SELLO ROJO MEDIO 250	\$3.00
SPAGHETTI DORIA CLASICA X	\$3.00
CHOCOL. DIANA 600Gx16PAG	\$2.00
FIDEO COMARRICO CLASICA X	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
+%	TAX: \$0.00

TOTAL: \$15.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024266 -001
DATE:11/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA
GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **ZZZ**

PURCHASE:

GTA. TOSH MIEL BS. 9X3	\$3.00
CAFE SELLO ROJO MEDIO 250	\$2.00
SPAGHETTI DORIA CLASICA X	\$4.00
GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
+%	TAX: \$0.00

TOTAL: \$15.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475023934 -001
DATE:27/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA
GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **PPP**

PURCHASE:

CAFE SELLO ROJO MEDIO 250	\$4.00
PASAB. LA ESPECIAL SAL AP	\$2.00
GTA. DUCALES TC. X6 720G	\$3.00
CHOCOL.CHOCOLYNE CLAS 6PL	\$5.00
JET WAFER VAINILLA 20PLEX	\$2.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024073 -001
DATE:24/09/2016 5:01:29 PM

Soporte de {Café Sello Rojo} = $3/4 = 75\%$

Soporte de {Spaghetti Doria Clásica} = $2/4 = 50\%$



REGLAS DE ASOCIACIÓN — ITEMSET FRECUENTE

¿Cuál es el soporte del itemset {Café Sello Rojo, Spaghetti Doria Clásica}?

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **XXX**

PURCHASE:

GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
JET WAFER VAINILLA 20PLEX	\$2.00
FIDEO COMARRICO CLASICA X	\$4.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024200 -001
DATE:21/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **YYY**

PURCHASE:

CAFE SELLO ROJO MEDIO 250	\$3.00
SPAGHETTI DORIA CLASICA X	\$3.00
CHOCOL. DIANA 500Gx16PAGT	\$2.00
FIDEO COMARRICO CLASICA X	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
+%	TAX: \$0.00

TOTAL: \$15.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024266 -001
DATE:11/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **ZZZ**

PURCHASE:

GTA. TOSH MIEL BS. 9X3	\$3.00
CAFE SELLO ROJO MEDIO 250	\$2.00
SPAGHETTI DORIA CLASICA X	\$4.00
GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
+%	TAX: \$0.00

TOTAL: \$15.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475023934 -001
DATE:27/09/2016 5:01:29 PM

DISFRUTAR LA VIDA TE ALIMENTA

GRUPO NUTRESA
CARRERA 52 NO. 20 - 124
MEDELLÍN
ANTIOQUIA
+57

CASHIER: 123
CUSTOMER: **PPP**

PURCHASE:

CAFE SELLO ROJO MEDIO 250	\$4.00
PASAB. LA ESPECIAL SAL AP	\$2.00
GTA. DUCALES TC. X6 720G	\$3.00
CHOCOL.CHOCOLYNE CLAS 6PL	\$5.00
JET WAFER VAINILLA 20PLEX	\$2.00
+%	TAX: \$0.00

TOTAL: \$16.00

PAYMENT METHOD: CREDIT CARD
TRANSACTION #1475024073 -001
DATE:24/09/2016 5:01:29 PM

Soporte de {Café Sello Rojo} = $3/4 = 75\%$

Soporte de {Spaghetti Doria Clásica} = $2/4 = 50\%$

Soporte de {Café Sello Rojo, Spaghetti Doria Clásica} = $2/4 = 50\%$



REGLAS DE ASOCIACIÓN — ITEMSET FRECUENTE

¿Confianza, lift y leverage de {Café Sello Rojo → Spaghetti Doria Clásica}?

DISFRUTAR LA VIDA TE ALIMENTA	
GRUPO NUTRESA CARRERA 52 NO. 20 - 124 MEDELLIN ANTIOQUIA +57	
CASHIER: 123	
CUSTOMER: XXX	
PURCHASE:	
GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
JET WAFER VAINILLA 20PLEX	\$2.00
FIDEO COMARRICO CLASICA X	\$4.00
+%	TAX: \$0.00
TOTAL: \$16.00	
PAYMENT METHOD: CREDIT CARD TRANSACTION #1475024200 -001 DATE:21/09/2016 5:01:29 PM	

DISFRUTAR LA VIDA TE ALIMENTA	
GRUPO NUTRESA CARRERA 52 NO. 20 - 124 MEDELLIN ANTIOQUIA +57	
CASHIER: 123	
CUSTOMER: YYY	
PURCHASE:	
CAFE SELLO ROJO MEDIO 250	\$3.00
SPAGHETTI DORIA CLASICA X	\$3.00
CHOCOL. DIANA 500Gx16PAGT	\$2.00
FIDEO COMARRICO CLASICA X	\$3.00
GTA. TOSH CHOCOLATE BS.	\$4.00
+%	TAX: \$0.00
TOTAL: \$15.00	
PAYMENT METHOD: CREDIT CARD TRANSACTION #1475024266 -001 DATE:11/09/2016 5:01:29 PM	

DISFRUTAR LA VIDA TE ALIMENTA	
GRUPO NUTRESA CARRERA 52 NO. 20 - 124 MEDELLIN ANTIOQUIA +57	
CASHIER: 123	
CUSTOMER: ZZZ	
PURCHASE:	
GTA. TOSH MIEL BS. 9X3	\$3.00
CAFE SELLO ROJO MEDIO 250	\$2.00
SPAGHETTI DORIA CLASICA X	\$4.00
GTA. SALTINNOEL ROJO TC.X	\$3.00
GTA. MINICHIPS CHOCOLATE	\$3.00
+%	TAX: \$0.00
TOTAL: \$15.00	
PAYMENT METHOD: CREDIT CARD TRANSACTION #1475023934 -001 DATE:27/09/2016 5:01:29 PM	

DISFRUTAR LA VIDA TE ALIMENTA	
GRUPO NUTRESA CARRERA 52 NO. 20 - 124 MEDELLIN ANTIOQUIA +57	
CASHIER: 123	
CUSTOMER: PPP	
PURCHASE:	
CAFE SELLO ROJO MEDIO 250	\$4.00
PASAB. LA ESPECIAL SAL AP	\$2.00
GTA. DUCALES TC. X6 720G	\$3.00
CHOCOL.CHOCOLYNE CLAS 6PL	\$5.00
JET WAFER VAINILLA 20PLEX	\$2.00
+%	TAX: \$0.00
TOTAL: \$16.00	
PAYMENT METHOD: CREDIT CARD TRANSACTION #1475024073 -001 DATE:24/09/2016 5:01:29 PM	

Soporte de {Café Sello Rojo} = $3/4 = 75\%$

Soporte de {Spaghetti Doria Clásica} = $2/4 = 50\%$

Soporte de {Café Sello Rojo, Spaghetti Doria Clásica} = $2/4 = 50\%$

Confianza de {Café Sello Rojo → Spaghetti Doria Clásica} = $(2/4)/(3/4) = 66,6\%$

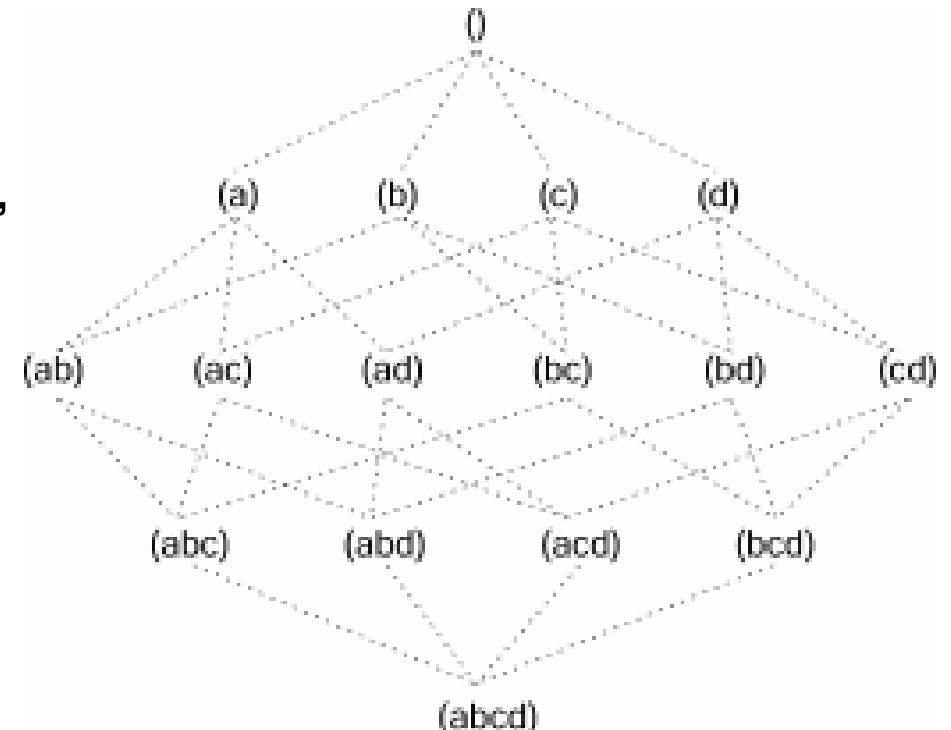
Lift de {Café Sello Rojo → Spaghetti Doria Clásica} = $(2/4) / ((3/4)*(2/4)) = 4/3 = 1,33$

Leverage de {Café Sello Rojo → Spaghetti Doria Clásica} = $(2/4) - ((3/4)*(2/4)) = 4/3 = 0,125$



APRIORI

- Encontrar los itemsets frecuentes es un problema Np-Hard.
- El algoritmo **Apriori** poda el espacio de búsqueda, para luego definir las reglas resultantes
- Búsqueda bottom-up de los itemsets frecuentes:
 - se debe especificar un umbral de **sopORTE** mínimo
- Las reglas son extraídas de los itemsets frecuentes encontrados:
 - se pueden especificar condiciones adicionales para las reglas encontradas con respecto a métricas de **confianza**, **lift** y/o **leverage**.
 - Itemset antecedente → Itemset consecuente

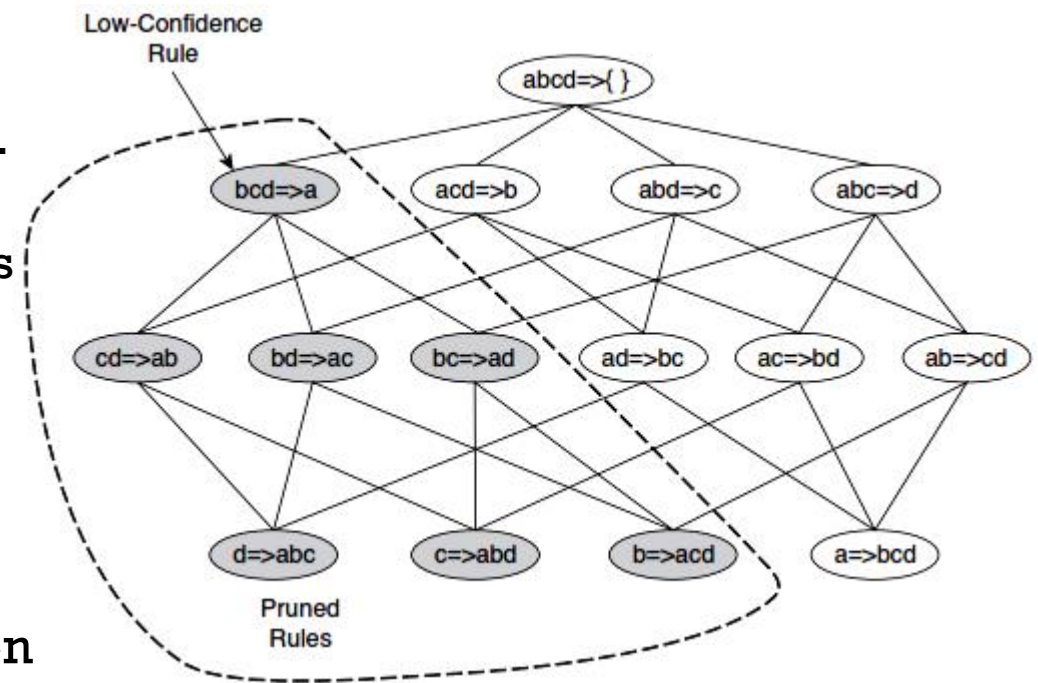


Espacio de búsqueda de 4 ítems



APRIORI - ALGORITMO

- Encontrar los itemsets candidatos (para un **soporte** definido):
 - Retener los 1-itemsets frecuentes como candidatos.
 - Descartar los 1-itemsets no frecuentes
 - Para (n en 2:N): Encontrar los (n)-itemset frecuentes combinando los (n-1)-itemsets candidatos. Guardarlos como candidatos
 - Repetir hasta el final, hasta que los itemsets se queden por debajo del soporte, o hasta llegar a un máximo de cardinalidad especificada
- Combinar los miembros de los itemsets candidatos, encontrando las reglas que satisfacen un mínimo de otra medida definida (**confidence**, **lift**, **leverage**)



Ejemplo de resultados encontrados

<http://www.paulallen.ca/apriori-algorithm-rule-generation/>



APRIORI - VALIDACIÓN

Las reglas encontradas por Apriori deben ser evaluadas

- Utilizar soporte, confianza, lift y leverage para definir interés y significancia de las reglas
 - La especificación de umbrales para las medidas es un proceso iterativo
- Basarse en argumentos subjetivos derivados de conocimiento del dominio de aplicación. Eliminar reglas obvias.
 - Carne de hamburguesa → ketchup
 - El 90% de las personas compran azúcar, eliminar reglas con consecuente azúcar
 - La sal no nos aporta nada en cuanto a utilidades, eliminar reglas con consecuente sal



APRIORI

- Consideraciones:
 - La búsqueda en anchura genera una **complejidad** computacional temporal y espacial alta: cuando el número de productos y/o transacciones es muy grande, es necesario adoptar estrategias adicionales para reducir el espacio de búsqueda
 - En grandes datasets, la mayoría de los eventos van a ser raros (soportes y confianzas bajas)
 - La minería de reglas de asociación debe hacerse iterativamente, teniendo en cuenta la opinión de expertos del dominio en el equipo de analítica
 - Las reglas obtenidas pueden de 3 tipos: triviales, inexplicables (aleatorias) y accionables
- Alternativas
 - Eclat: algoritmo de búsqueda en profundidad
 - FP Growth



TALLER: APRIORI (A MANO)

- **Taller:**
- Determinar los itemsets **frecuentes** de las transacciones siguientes, dado un umbral de soporte mínimo de 50%:
 - 1: A,B,C,E ▪ 4: A,C,D,E
 - 2: A,C,D,E ▪ 5: C,D,E
 - 3: B,C,E ▪ 6: A,D,E
- Determinar las reglas $X \rightarrow Y$, no filtradas por el umbral del 50%, que tengan al menos una confianza del 66% y que detecten productos complementarios que no lo sean por coincidencia.



TALLER: REGLAS DE ASOCIACIÓN

1. Cargar el dataset “supermarket.arff” en Weka; explorarlo.
2. En la pestaña “Associate” escoger el algoritmo “Apriori”.
 - a) Se define una métrica de evaluación de las reglas (metricType, e.g. “Confidence” por defecto) con un umbral mínimo a superar (minMetric=0,9), y el número de reglas que se quiere (numRules=10)
 - b) La implementación de Weka es iterativa haciendo intentos con diferentes niveles de soporte de los itemsets de base para la producción de reglas. Empieza con un nivel de soporte (upperBoundMinSupport=1), y trata de encontrar las reglas especificadas. Si no alcanza, disminuye el nivel de soporte (delta=0,05) y vuelve a intentar.
3. Lanzar el algoritmo. Weka reduce el soporte mínimo hasta 0,15 para encontrar las 10 reglas con confianza superior a 0,9.
4. Analizar las reglas



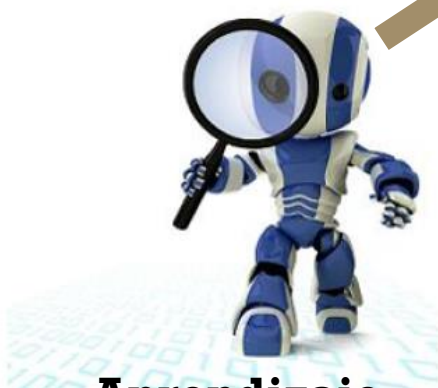
AGENDA



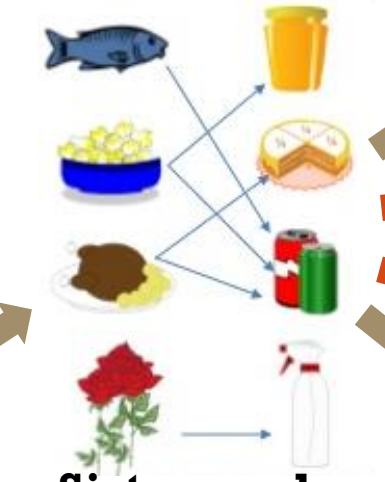
**Aprendizaje
automático**



**Aprendizaje
supervisado**



**Aprendizaje
no supervisado**



**Sistemas de
recomendación**

$$\text{tfidf}_{i,j} = \text{tf}_{i,j} \times \log \left(\frac{N}{\text{df}_i} \right)$$

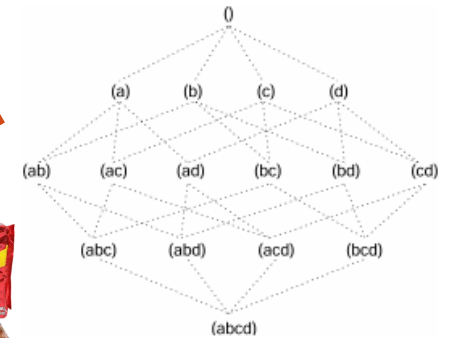
TF/IDF



Basados en contenido



Itemset



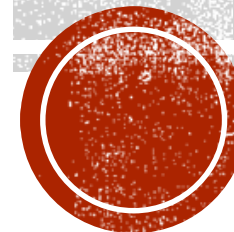
Apriori



BASADOS EN CONTENIDO



Recommendations for you in Music



BASADOS EN CONTENIDO

- **Sistemas basados en casos y en conocimiento**

- Subtipos de sistemas de recomendación
- Sistemas basados en queries
- Entrada de preferencias a través de entrevistas o scripts
- Preferencias de corta vida (noticias, hoteles, vuelos), modelos de usuarios sin persistencia

- **Sistemas de recomendación basados en contenido**

- Modelos de usuario con las preferencias con persistencia
- Proveen un buen mecanismo para encontrar productos sustitutos, pero no complementarios
- Preferencias de larga vida
- Fáciles de explicar a los usuarios
- Proveen una buena base para organizar la navegación de la base de ítems antes de la compra
- No tienen problema de arranque en seco (cold start)



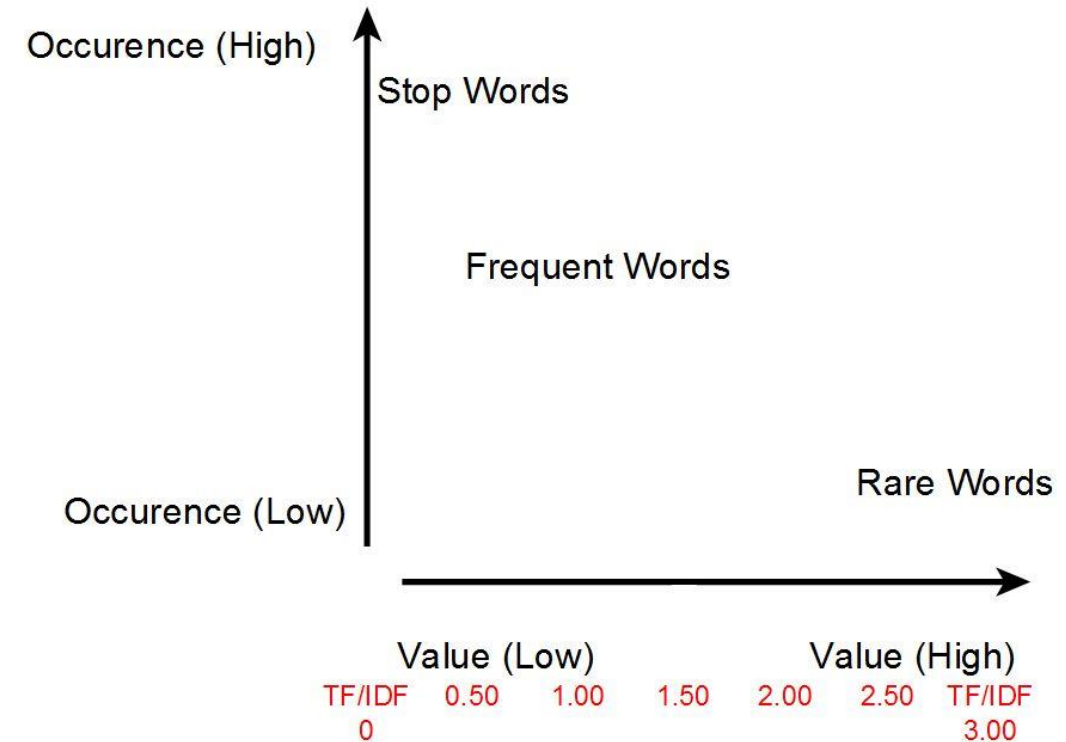
BASADOS EN CONTENIDO

- **Supuesto:** las preferencias de los usuarios son estables en el tiempo
- Un vector de atributos describe tanto los ítems como los usuarios (palabras clave, categorías, etiquetas, gustos)
- No se utiliza información de otros usuarios
- Modelo de usuario tiene en cuenta:
 - Modificaciones directas de usuario sobre su propio modelo
 - Interacciones explícitas (ratings, likes)
 - Interacciones implícitas (lecturas, clicks, compras)
- Descriptores de atributos en el vector:
 - Binarios
 - Conteos
 - TF/IDF



TF/IDF

- **TFIDF** (Term Frequency Inverse Document Frequency): Esquema de pesos usado para describir un ítem (documento) a partir de un vector de etiquetas (término)
 - Filtraje de documentos, motores de búsqueda, information retrieval
 - $TFIDF = TF * IDF$
- **TF**: Term Frequency
 - TF= #ocurrencias de un término en un documento
 - Se utilizar la transformación **log (TF +1)**, en caso de distribuciones alargadas
 - Para documentos de tamaños diferentes se puede adicionalmente normalizar con respecto al número de términos totales: **log (TF +1)/n**
- **IDF**: Inverse Document Frequency
 - $IDF = \log(\#documentos / \#documentos \text{ con el término})$
 - Entre más raro es un término, mayor su IDF

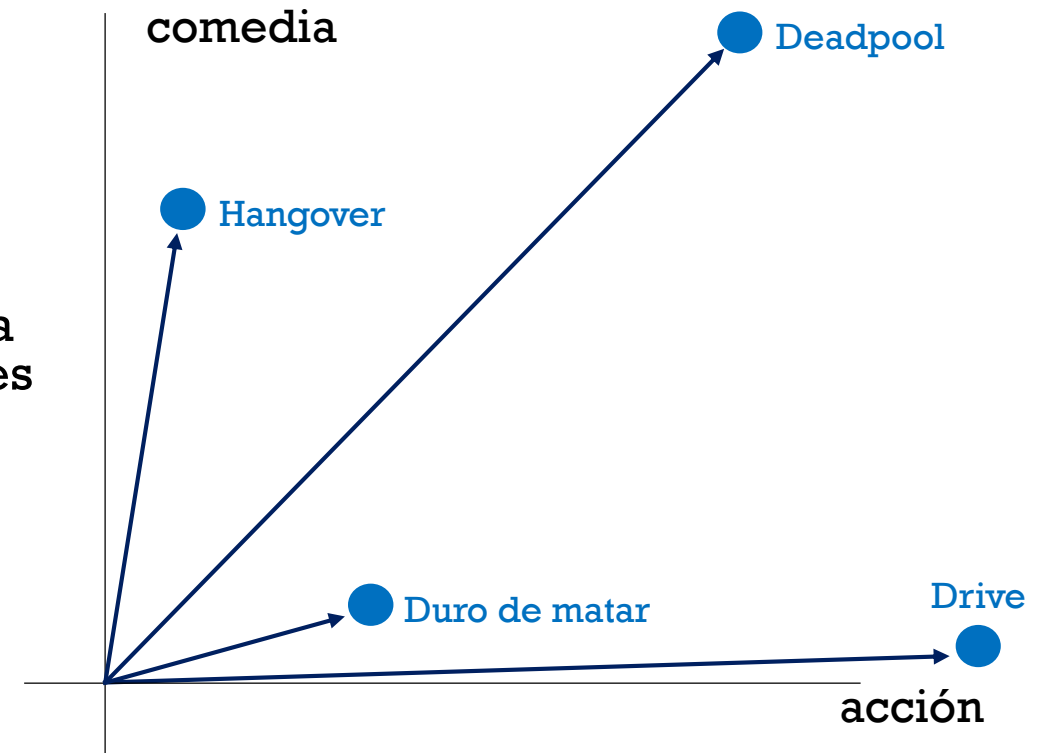


<http://trimc-nlp.blogspot.com.co/2013/04/tfidf-with-google-n-grams-and-pos-tags.html>



TF/IDF

- Cada usuario y cada ítem se representan como **vectores multidimensionales** en un espacio dado por los atributos de descripción de contenido
- Se debe **normalizar** (vectores de largo 1) ? Pérdida de información vs. ítems con importancias diferentes

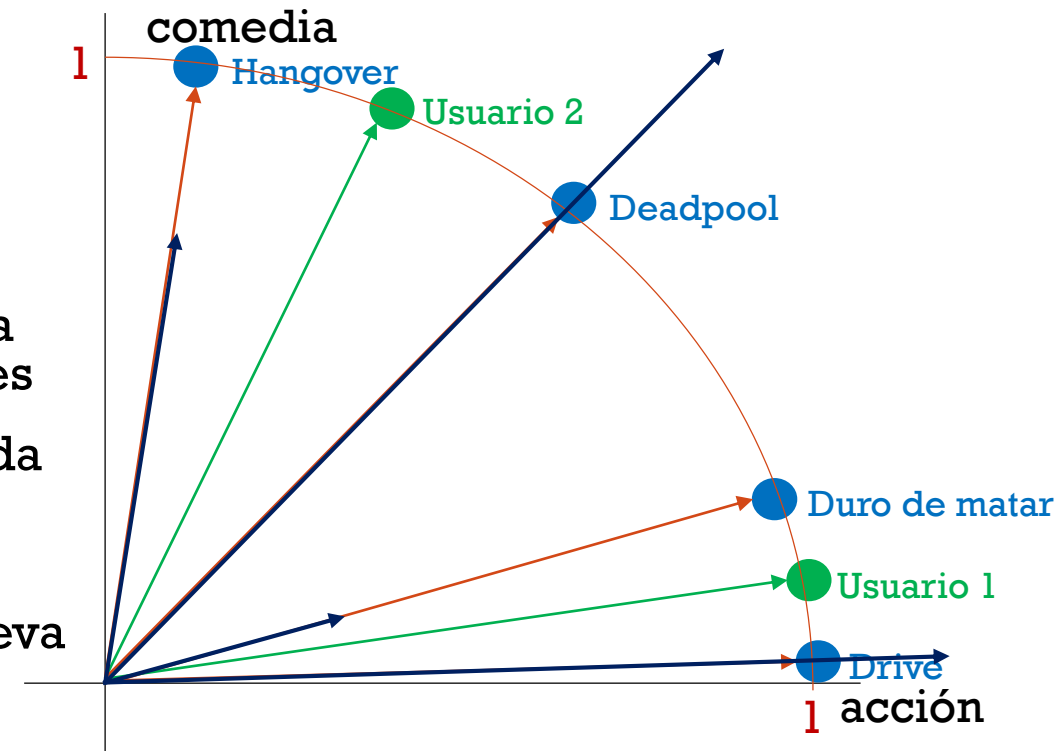


$$\|\mathbf{q}\| = \sqrt{\sum_{i=1}^n q_i^2}$$



TF/IDF

- Cada usuario y cada ítem se representan como **vectores multidimensionales** en un espacio dado por los atributos de descripción de contenido
- Se debe **normalizar** (vectores de largo 1) ? Pérdida de información vs. ítems con importancias diferentes
- **Evidencia** de una preferencia puede ser recolectada explícitamente o implícitamente
- **Modelos de usuario**: Construidos desde cero o actualizados después de consideración de cada nueva evidencia (positive o negativa), después de cada interacción del usuario
- **Predicción** de preferencia: similitud entre un usuario y un ítem



$$\|\mathbf{q}\| = \sqrt{\sum_{i=1}^n q_i^2}$$



TF/IDF

■ Medidas de **similitud** a considerar

■ Producto interno:

- Sin límites
- Sensible a la magnitud de cada vector

$$Inner(x, y) = \sum_i x_i y_i = \langle x, y \rangle$$

■ Coseno:

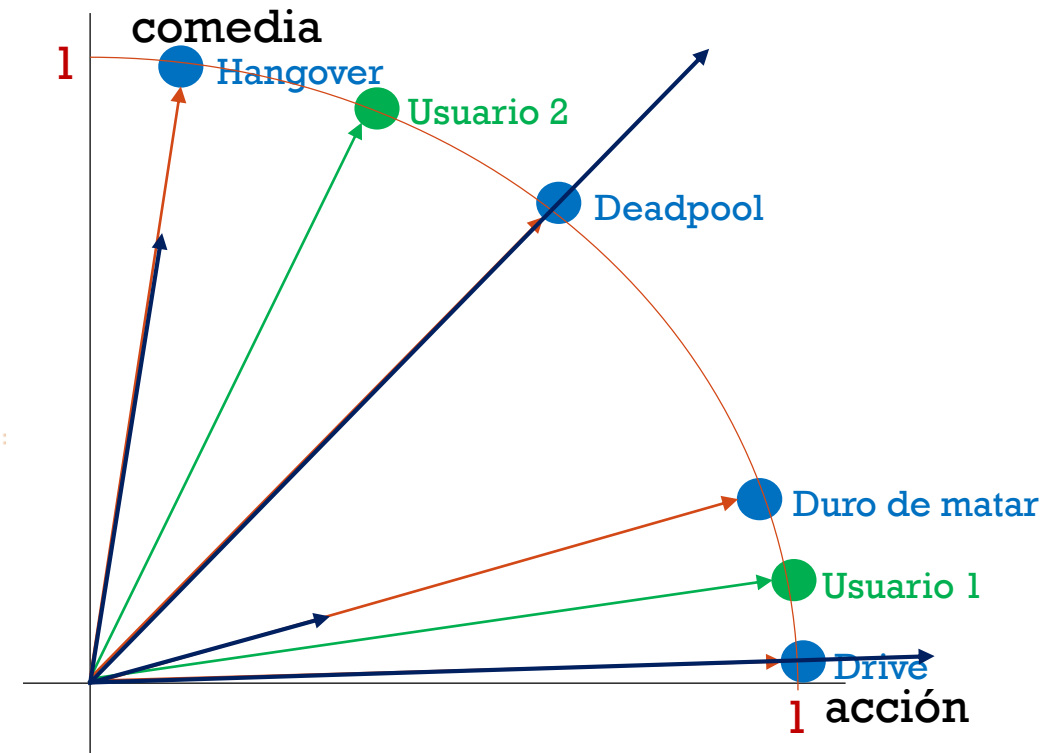
- Límites [-1; +1]
- Normalización según las magnitudes
- No es invariante a translaciones (posición)

$$CosSim(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \sqrt{\sum_i y_i^2}}$$

■ Correlación de Pearson

- Límites [-1; +1]
- Equivalente al coseno entre versiones centradas de los vectores comparados
- Invariante en cuanto a escala y posición

$$Corr(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$



$$\| \mathbf{q} \| = \sqrt{\sum_{i=1}^n q_i^2}$$



TF/IDF

- PRO: Modelo para todo tipo de sistemas de recomendación basados en contenido
- CONTRA:
 - Número de dimensiones puede convertirse en un problema
 - No considerar los “stop words”
 - Utilizar agrupamientos de términos en categorías (persecución, pistolas, policíacas, guerra → acción)
 - Reducción de dimensionalidad
 - PCA
 - Análisis de semántica latente
 - Problemas con eufemismos (e.g. contratos)
 - No consideración de frases o n-gramas
 - No consideración de adyacencia ni orden (bag of words)
 - No consideración de la interdependencia (Me gustan las películas serias de R.De Niro pero no sus comedias, me gustan las comedias de Woody Allen pero no sus películas serias)
 - No consideración de la importancia contextual de los términos (títulos, descripción, cuerpo)
 - No consideración de contenidos implícitos
 - No consideración de los cambios de gustos → se puede solventar con pesos



TALLER: BASADO EN CONTENIDO (EXCEL)

Descargar:

- El enunciado del taller de sistemas de recomendación basado en contenido en Excel,
- Hoja de Excel con los datos de trabajo.

Desarrollar el taller.



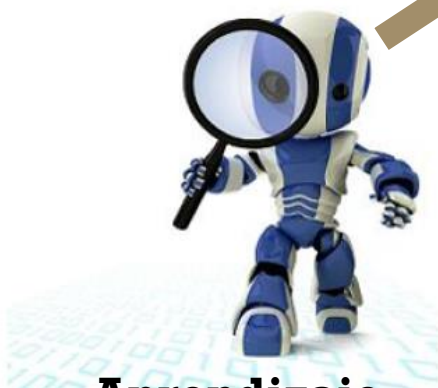
AGENDA



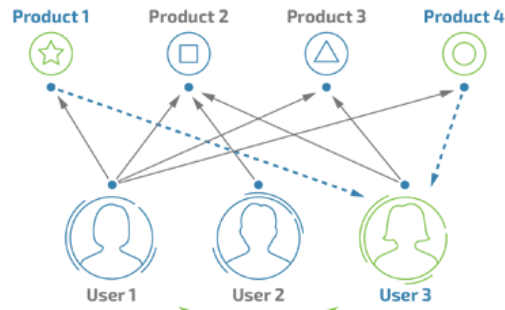
Aprendizaje automático



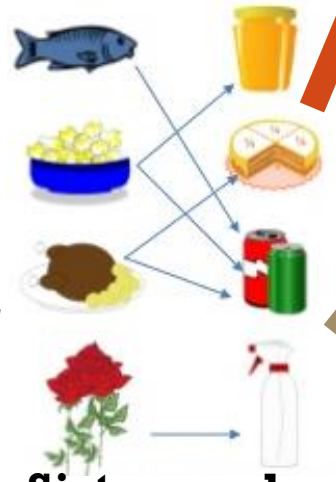
Aprendizaje supervisado



Aprendizaje no supervisado



Filtro colaborativo



Sistemas de recomendación

$$tfidf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

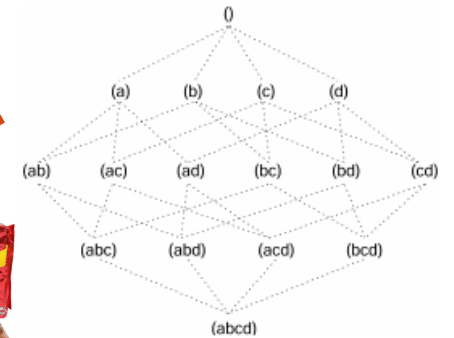
TF/IDF



Basados en contenido



Itemset



Apriori



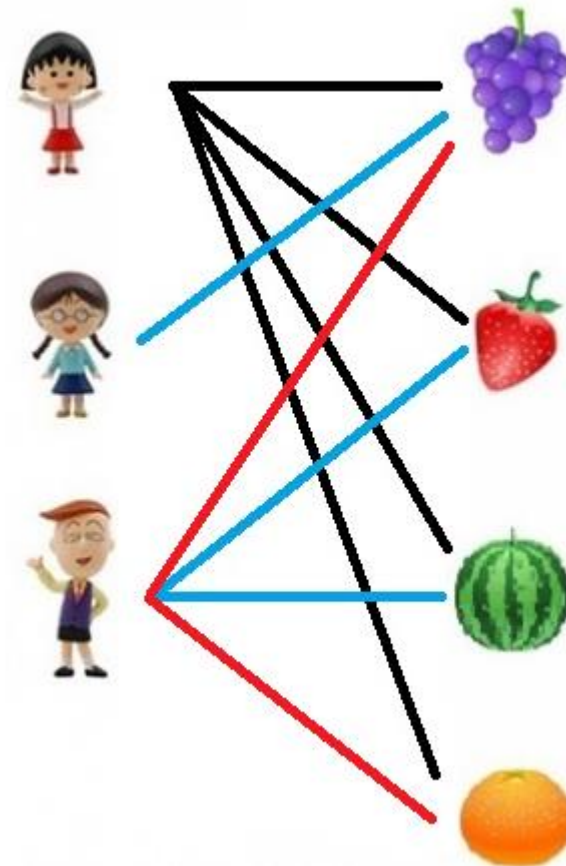
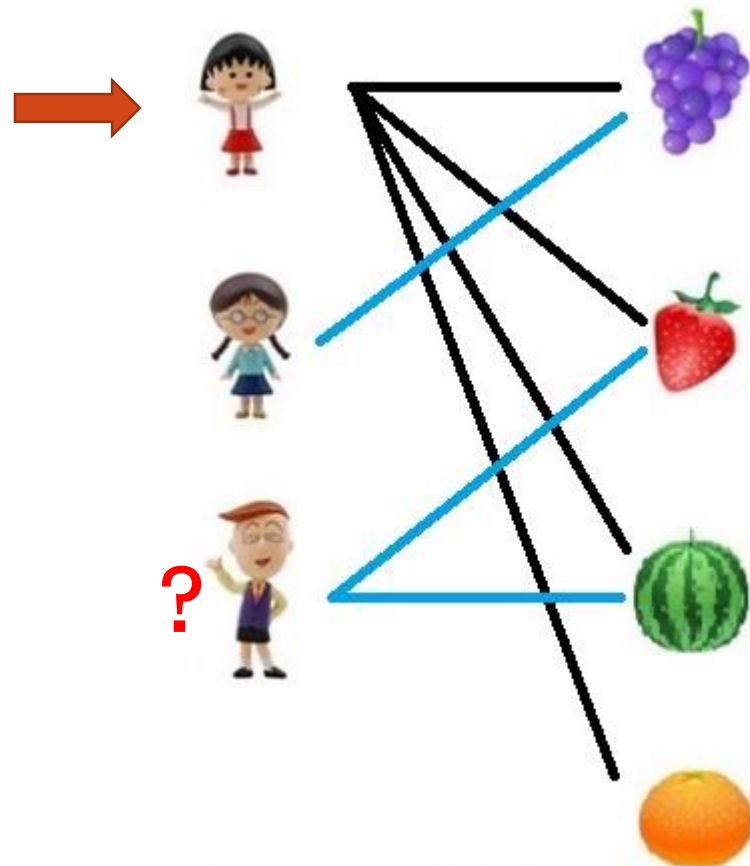
FILTRO COLABORATIVO

Suposiciones:

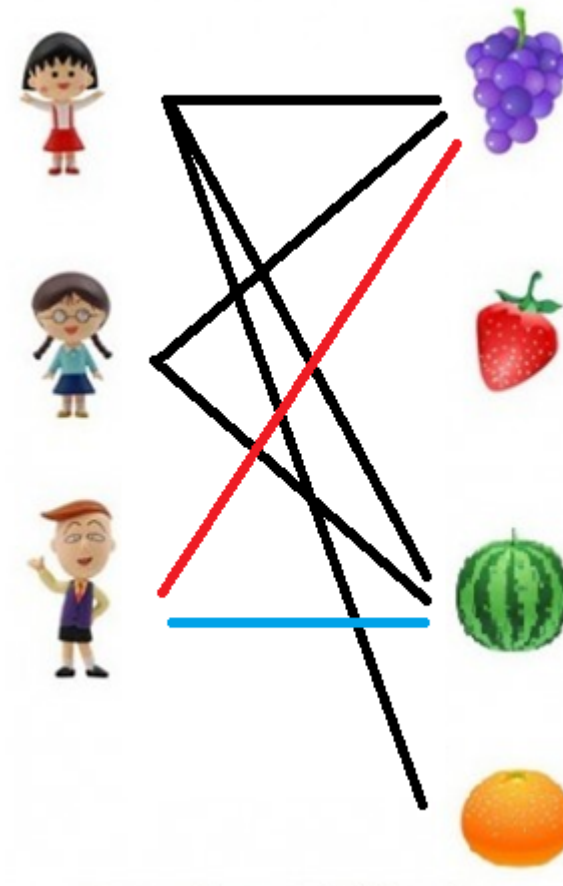
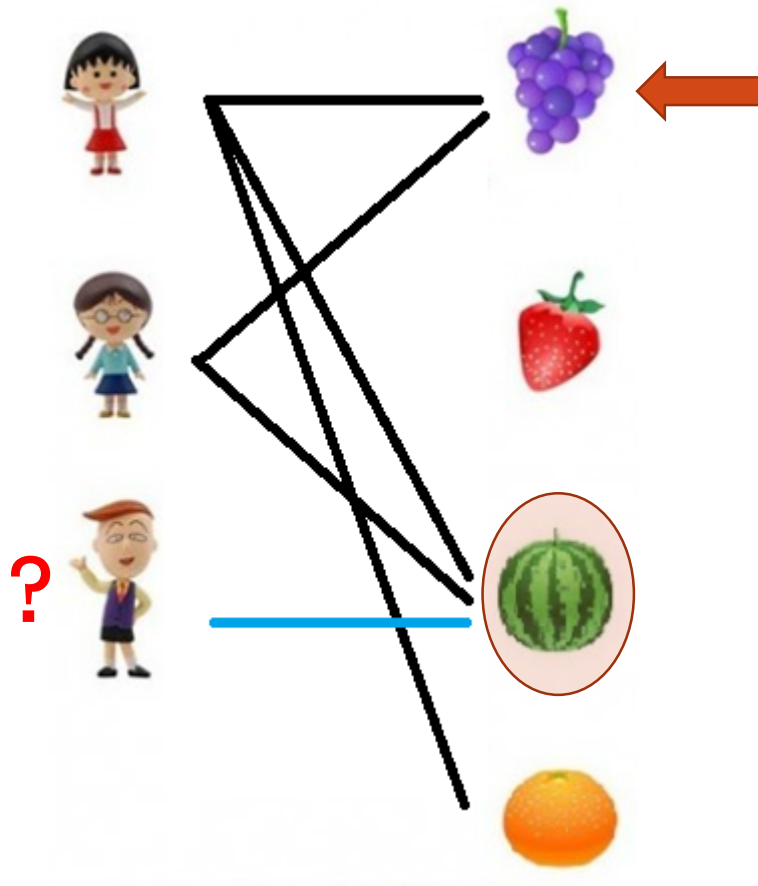
- El comportamiento pasado permite predecir el comportamiento futuro
- Los gustos son estables o se mueven de manera sincronizada para todos los usuarios
- Hay acuerdo en el contexto del dominio de los ítems que se han notado y que se recomendarán (usuarios pueden estar de acuerdo en películas, pero no necesariamente en política, humor, hoteles y restaurantes)



FILTRO COLABORATIVO – BASADO EN USUARIO



FILTRO COLABORATIVO – BASADO EN ÍTEMS



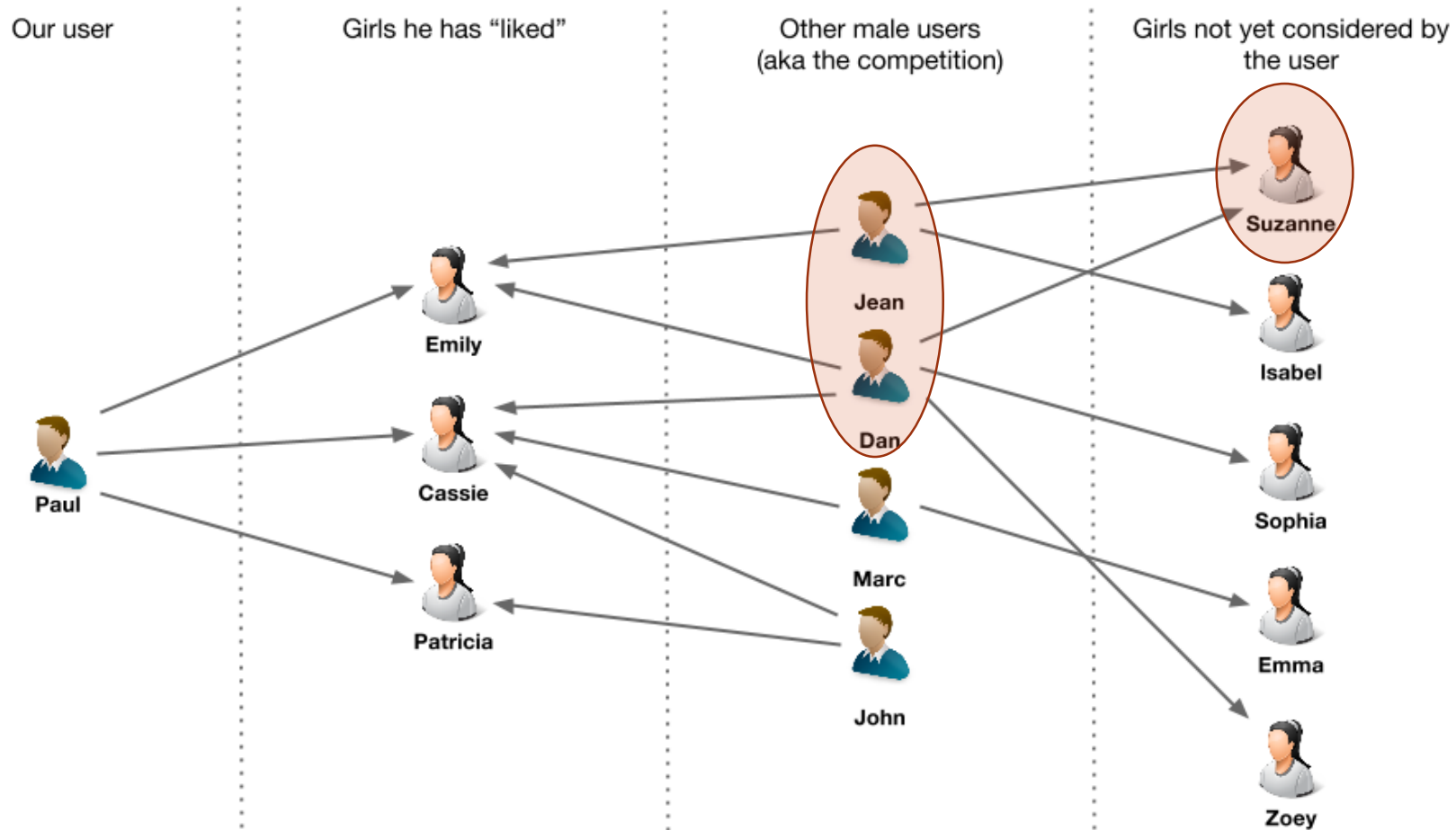
FILTRO COLABORATIVO

- Matriz Usuario X Ítem
- Significados de ratings diferentes
 - Distribución
 - Normalización
 - Por usuario? Por ítem?
- Distancias & similitudes
- Filtro colaborativo basado en usuarios
 - Juan Pablo vs. Daniel
- Filtro Colaborativo basado en ítems
 - Drive vs. Inception

Pelicula																		
Usuarios	Drive	Inception	Star Wars	Avengers	Suicide Squad	X-Men Apocalypse	Conjuring 2	10, Cloverfield Lane	Mad Max: Fury Road	The Martian	The Revenant	Spotlight	The Danish Girl	Room	Zootopia	Frozen	El Bebé de Bridget Jon	Twilight
Johan	4	4	5	5	5	5	3	3	5	3	1	1	1	1	2	1	1	1
Andres	5	5	4	3	2	3	5	5	5	4	4	4	4	4	2	2	1	1
Melisa	3	3	4	2	2	3	3	2	1	5	3	4	5	5	5	5	5	5
Nicholle	3		4	5	3	4	5	2	1	4	3	5	5	4	5	5	5	5
Juan Pablo	3	3	2	3	1	2	2	3	3	2	1	2	1	1	1	1	1	
Daniel	5	5	4	5	4	4	4	5	5	4	3	4	3	3	3	4	3	5



FILTRO COLABORATIVO — BASADO EN USUARIOS



<https://linkurio.us/using-neo4j-to-build-a-recommendation-engine-based-on-collaborative-filtering/>



FILTRO COLABORATIVO — BASADO EN USUARIOS

- Algoritmo Usuario-Usuario
 - Calcular la matriz de similitud entre usuarios
 - Para un usuario dado, definir su “barrio” de usuarios parecidos (K-NN)
 - Calcular una medida de recomendación para los ítems relacionados (promedio, promedio ponderado, regresión lineal múltiple)
- Características:
 - No depende del contenido (descripción de los ítems), fuente de errores
 - Puede aplicarse a todo tipo de dominios (no es el caso del basado en contenido)
 - Recomendaciones casuales (serendipia)
- Varias consideraciones a tener en cuenta



FILTRO COLABORATIVO — BASADO EN USUARIOS

Consideraciones:

- Selección del conjunto próximo de usuarios:
 - Todos los usuarios
 - Al azar
 - Utilizar un umbral de similitud/distancia
 - Top-N usuarios según similitud/distancia
 - Top-N usuarios según similitud/distancia con respecto al ítem en consideración
 - Puede haber demasiados usuarios en la proximidad definida
 - Puede no haber ningún usuario en la proximidad definida
- Cobertura de recomendación



FILTRO COLABORATIVO — BASADO EN USUARIOS

Consideraciones:

- Normalización
 - Hay usuarios que les gusta todo, otros a los que no les gusta nada
 - Hay usuario que utilizan una escala de notación más amplia que otros
 - Promediar ignora estas diferencias
 - La normalización las tiene en cuenta



FILTRO COLABORATIVO — BASADO EN USUARIOS

Consideraciones:

■ Formulación de predicciones:

- Sin personalización:

$$P_{a,i} = \frac{\sum_{u=1}^n r_{u,i}}{n}$$

- Grado de similitud de usuarios:

$$P_{a,i} = \frac{\sum_{u=1}^n r_{u,i} * w_{a,u}}{\sum_{u=1}^n w_{a,u}},$$

$w_{a,u}$ puede ser una medida de similitud o de pertenencia al conjunto de usuarios próximos

- Normalización de ratings:

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^n (r_{u,i} - \bar{r}_u) * w_{a,u}}{\sum_{u=1}^n w_{a,u}}$$

- Estandarización:

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^n z_{u,i} * w_{a,u}}{\sum_{u=1}^n w_{a,u}} * \sigma_a$$

Pero, ratings predichos pueden salirse del rango → floor/ceiling



FILTRO COLABORATIVO — BASADO EN USUARIOS

Consideraciones:

- Medida de acuerdo entre usuarios
 - Solo se consideran los ratings de los ítems que ambos usuarios han notado
 - Correlación de Pearson : $w_{a,u} = \frac{\sum_{i=1}^m (r_{a,i} - \bar{r}_a) * (r_{u,i} - \bar{r}_u)}{\sigma_u * \sigma_a}$
 - Impráctica con datos unarios,
 - Impráctica para un número pequeño de ítems mutuamente anotados
 - Coseno: $w_{a,u} = \frac{\sum_{i=1}^m (r_{a,i} * r_{u,i})}{\|r_a\| * \|r_u\|}$
 - Cada usuario representado por un vector de sus ratings
 - Usada para datos binarios
 - Si normalizamos los ratings por usuario, se convierte en la correlación de Pearson
 - No considerar usuarios con muy pocos ratings en común (sesgo)



FILTRO COLABORATIVO — BASADO EN USUARIOS

Consideraciones:

- Información insuficiente:
 - Regularización (agregar K al denominador de la medida de similitud)
- Dimensionalidad (millones de usuarios (m), millones de ítems (n))
 - Creación matriz de similitud es $O(m^2 * n)$
 - Recomendación es $O(m*n)$
- Problema de arranque en frío (“Cold start”)
 - No hay datos para generar predicciones o recomendaciones
- Particionamiento de usuarios en clusters (bajos resultados)
- Pre-calcular vs. cambios en preferencias



FILTRO COLABORATIVO — BASADO EN ÍTEMS

- Proceso análogo al del filtro colaborativo basado en usuarios
 - El proceso no se hace fila por fila sino columna por columna en la matriz de usuarios (filas) por ítems (columnas)
- Algoritmo Ítem-Ítem
 - Calcular la matriz de similitud entre ítems
 - Para un ítem dado, definir su “barrio” de ítems parecidos (K-NN)
 - Calcular una medida de recomendación para los ítems relacionados



FILTRO COLABORATIVO — BASADO EN ÍTEMS

- Por qué?:
 - Complejidad de computación:
 - Consumo de recursos (CPU, memoria)
 - Las matrices no pueden ser pre calculadas
 - Cambios constantes
 - Si Usuarios >> ítems
 - Relaciones entre ítems es estable,
 - Muchos ratings por ítem (menos dispersos, facilidad de cálculo de similitudes)
 - Ítems son mas estables que usuarios
 - Pero
 - Complejidad de los usuarios no puede ser detectada por Ítem vs Ítem



TALLER: FILTRO COLABORATIVO BASADO EN USUARIOS EN EXCEL

Descargar el taller de sistemas de recomendación por filtro colaborativo y desarrollarlo.



REFERENCIAS

- *Introduction to recommender Systems*, Joseph Konstan, 2015
- EMC2, “Data science and big data analytics”, 2015, John Wiley & Sons
- *Data Science for Business*, Foster Provost & Tom Fawcett, O’Reilly, 2013
- *Practical Data Science with R*, Nina Zumel & John Mount, 2014
- *Mining association rules between sets of items in large databases*, R. Agrawal, T. Imielinski, and A. Swami, en Proc. of SIGMOD'93, 2013
- *Discovering frequent closed itemsets for association rules*, N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, en Proc. of ICDT'99, 1999
- http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp

