SISTEMAS DE RECOMENDACIÓN

JESUS DANIEL HERNANDEZ LONDOÑO
DEPARTAMENTO DE MATEMÁTICAS
UPR - MAYAGÜEZ

Motivación

- Amazon 45 TB de información de 60 millones de clientes.
- Hasta el 2016,Google busca en mas de 130 trillones de paginas, que representa mas de 390 Petabytes.
- Hay la necesidad de convertir datos en conocimiento e información.
- Ofrecer una propuesta de contenido a cada cliente será el factor diferencial para las empresas en ser exitosas o ignoradas por el consumidor (Mata, 2015).
- Netflix prize”

Introducción

Los sistemas de recomendación buscan calificar la preferencia de un usuario sobre algún ítem, basados en la interacción previa entre usuarios y elementos, porque los intereses y las tendencias pasadas son a menudo buenos indicadores de elecciones futuras.

Surgieron a mediados de la década de los 90, dando recomendaciones a los usuarios sin personalización.

Basados en contenido “Muéstrame más de lo similar a lo que me ha gustado”, Filtrado Colaborativos “Dime qué es popular entre mis compañeros”, Basados en el Conocimiento “Dime que se ajusta a mis necesidades”, e Híbridos (mezcla de los anteriores).

Sistemas de Recomendación

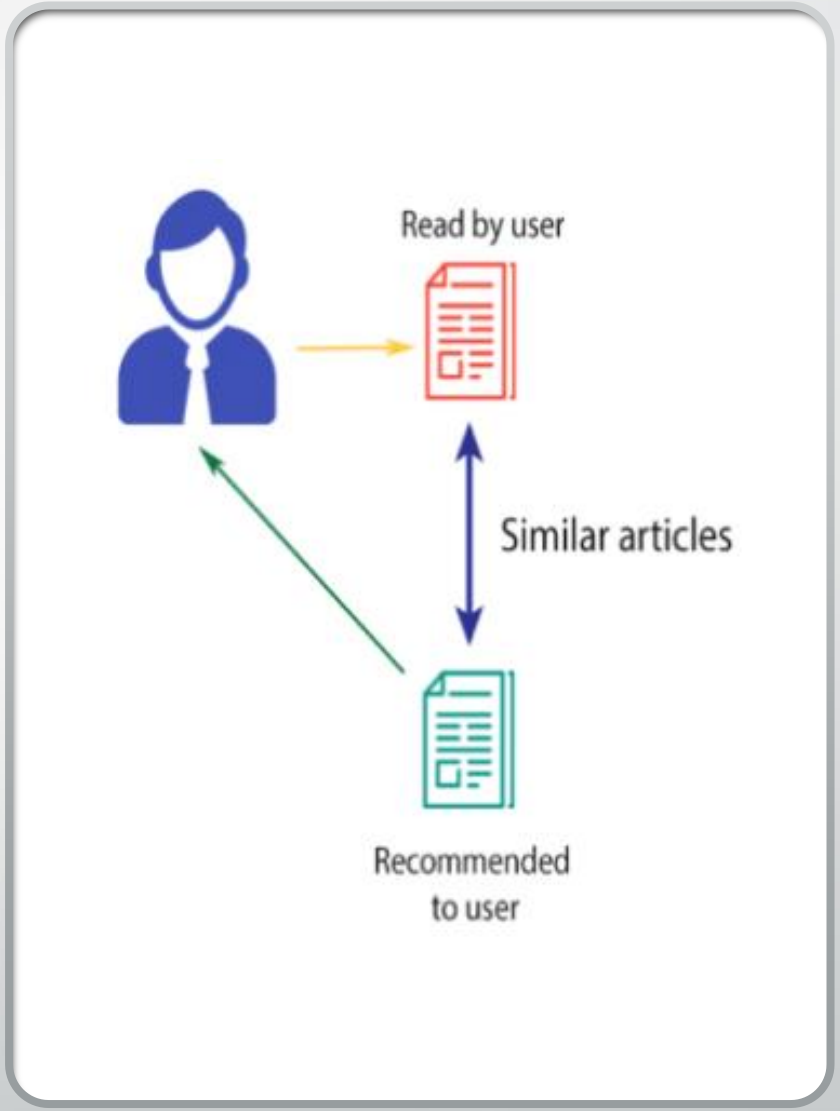
- Los sistemas de recomendación son técnicas, herramientas o algoritmos, que proporcionan al usuario información sobre elementos que sean de su interés (películas, libros, productos, entre otras cosas dependiendo de la compañía) de manera automatizada.

Basados en Contenido (content based filtering)

"Muéstrame más de lo similar a lo que me ha gustado"

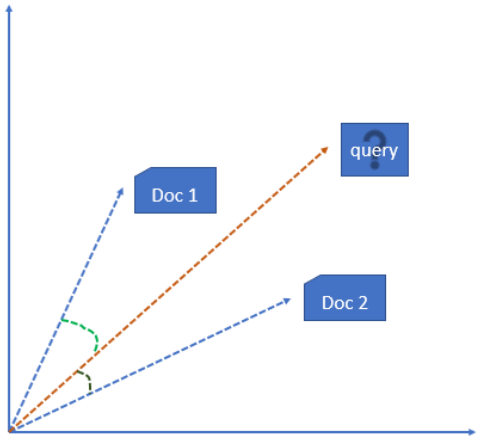
Limitaciones

- Se recomiendan productos similares a los ya consumidos por el usuario, poca originalidad.
- Los atributos/características con las que se describe los productos no aportan información acerca de la calidad del producto.



Algoritmos

- Modelo de espacios vectoriales basados en palabras clave.
- Naïve Bayes.
- TF-IDF o Frecuencia de documento - Frecuencia inversa de término: Asigna la importancia a una palabra en función del número de veces que aparece en el documento.



$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

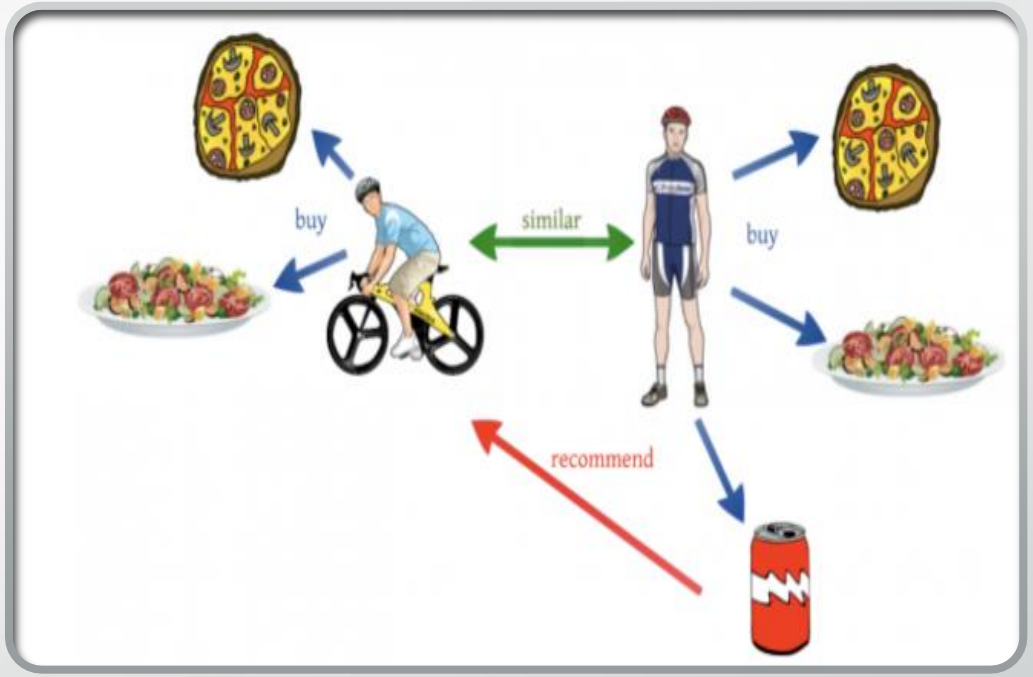
N = total number of documents

Filtrado colaborativo (collaborative filtering)

”Dime qué es popular entre mis compañeros”

Limitaciones

- No se puede hacer recomendaciones a nuevos usuarios.
- Los gustos de los usuarios pueden cambiar en el tiempo.



Métodos basados en Memoria

Basados en los vecinos.

- Usuario - Usuario (User - User): usuarios “similares” y recomendar algo que el usuario no ha visto.
- Ítem – ítem: similaridad entre artículos en función de las valoraciones que ha recibido.

Métodos basados en Modelos

- Árboles de decisión, modelos basados en reglas, métodos bayesianos y modelos de factores latentes, estos últimos tienen un alto nivel de cobertura.
- Precisión del modelo: RMSE

$$RMSE = \sqrt{\frac{\sum_{(u,j)\in E} e_{uj}^2}{|E|}}$$

Factores Latentes

- Modelos que aprovechan métodos de reducción de dimensionalidad, rotando el sistema de ejes, de modo que se eliminen las correlaciones, por ejemplo, Análisis de componentes principales (PCA) y la Descomposición de valores singulares (SVD).

Descomposición de valores singulares (SVD)

- Q_k y P_k, son ortogonales y contiene respectivamente los k vectores propios más grandes de RR^T y $R^T R$. La matriz Σ_k contiene las raíces cuadradas (no negativas) de los k valores propios más grandes de la matriz R a lo largo de su diagonal.

$$R \approx Q_k \Sigma_k P_k^T$$

$$U = Q_k \Sigma_k$$
$$V = P_k$$

$$R = UV^T$$

Minimize $J = \frac{1}{2}||R - UV^T||^2$

subject to:

Columns of U are mutually orthogonal

Columns of V are mutually orthogonal

Matriz de Votos (N x M)

	3	7	9	•	•
	•	•	8	3	7
	9	3	•	•	8
	7	9	2	3	6

Factores de los Usuarios (F X N)

F1	-0,4	0	0,8	0,4
F2	0,4	0	-0,4	0,8
F3	0,8	0,6	0	-0,8

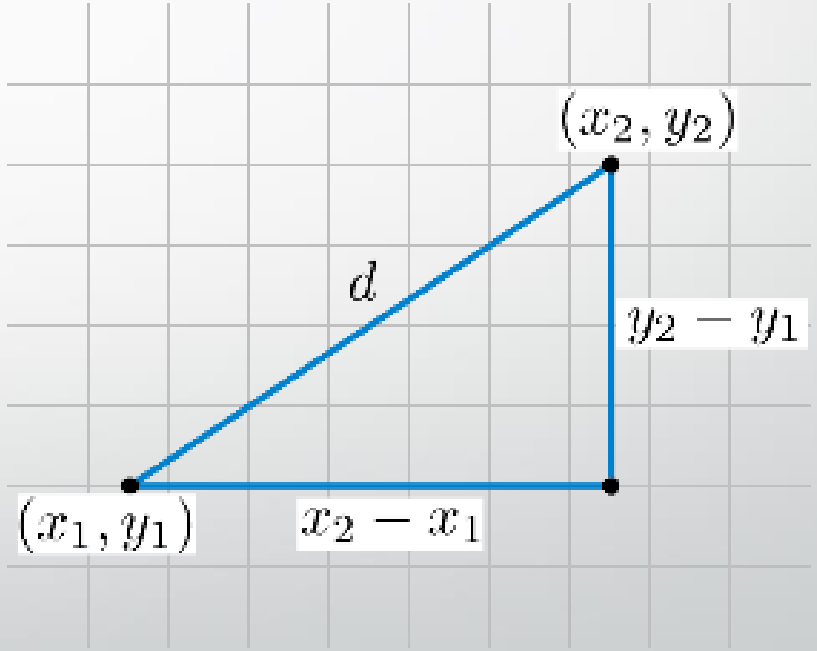
Factores de los Items (F X M)

F1	1	0,2	-0,8	0,6	0
F2	-0,8	1	-0.8	0.6	0,2
F3	-0,8	-0,2	1	0,2	0

Medidas de Similitud

Distancia Euclidiana

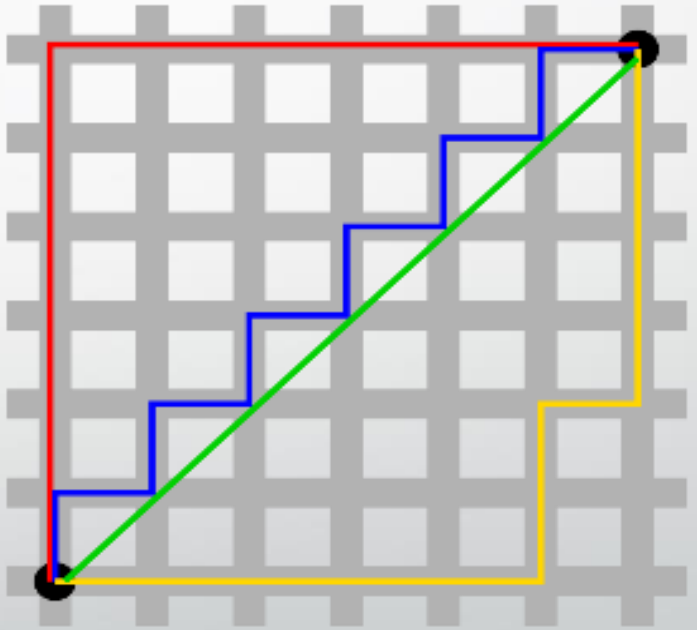
$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



Distancia Manhattan

$$d_{man}(p, q) = \sum_{i=1}^{n} |(p_i - q_i)|$$

$$|x_1 - u_1| + |x_2 - u_2| + \cdots + |x_n - u_n|$$



Correlación

- Medida que estudia la relación lineal.
- Pearson, Spearman, Kendall.
- Pearson asume distribución normal, sin embargo, se menciona que sigue siendo bastante robusta si no se considera normalidad, sensible a los valores extremos.
- Spearman es un método no paramétrico, datos ordinales, de intervalo.
- Kendall es un método no paramétrico, trabaja con rangos. Se emplea cuando se dispone de pocos datos.

Correlación de Pearson

$$\rho = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

$$\text{cov}(X,Y) = E(XY) - (EX)(EY)$$

Coeficiente de coincidencia simple (Simple Matching Coefficient (SMC))

- M_ {01} y M_ {10} son el número de variables que no coinciden y M_ {11} y M_ {00} el número de variables para las que ambas observaciones tienen el mismo valor.
- 1-SMC

		A	
		0	1
B	0	M_{00}	M_{10}
	1	M_{01}	M_{11}

$$SMC = \frac{\text{número coincidencias}}{\text{número total de atributos}} = \frac{M_{00} + M11}{M_{00} + M_{01} + M_{10} + M_{11}}$$

Índice Jaccard

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

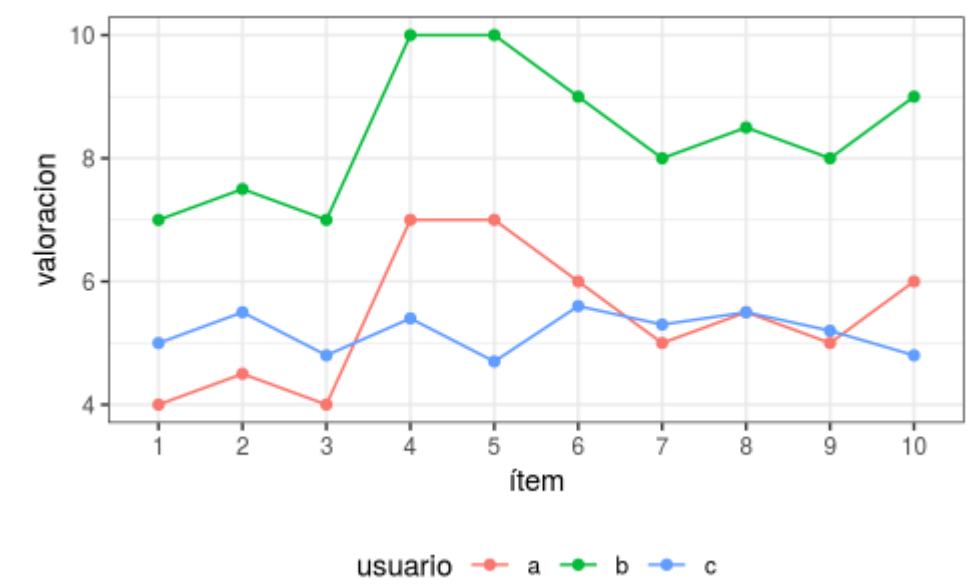
- Supermercado con 1000 productos y dos clientes. La canasta del primer cliente contiene sal y pimienta y la canasta del segundo contiene sal y azúcar. En este escenario, el índice Jaccard sería 1/3, pero usando el SMC se convierte en 0.998, definiendo M_ {11} como el número de coincidencia entre las dos canastas y M_ {00} el complemento de coincidencia.

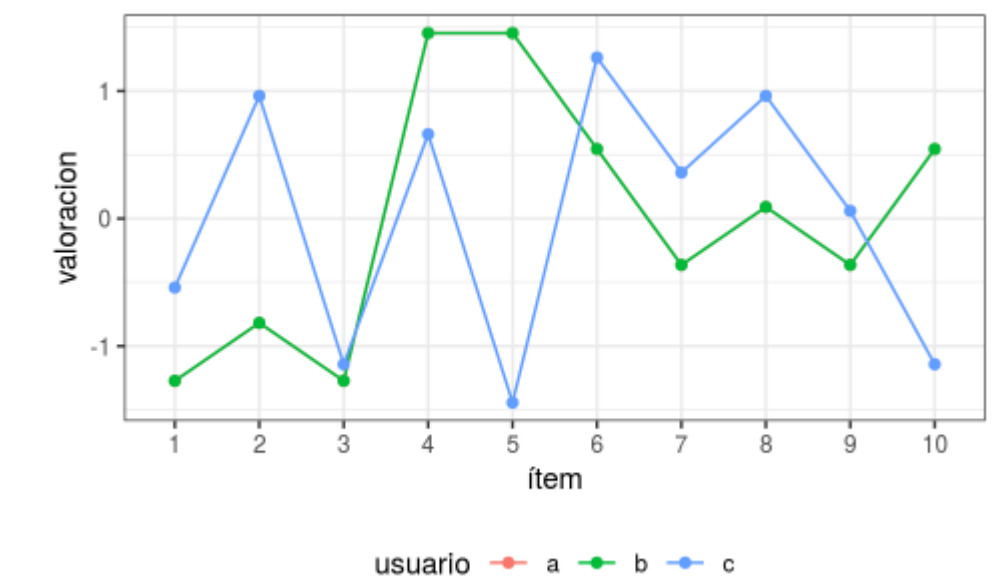
Similitud del Coseno

- Si el ángulo es 0° es decir los dos vectores son paralelos su coseno es 1.
- Si el ángulo es 90° es decir los dos vectores son perpendiculares su coseno es 0.
- Si el ángulo es de 180° es decir los dos vectores van en sentido contrario su coseno es -1.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Estandarización de valoraciones





Referencias

- Charu C. Aggarwal, *Recommender Systems.* IBM T.J. Watson Research Center Yorktown Heights, NY, USA
- Terveen, L., & Hill, W., (2001). Beyond Recommender Systems: Helping People Help Each Other. *HCI in the New Millennium*, 1, pp. 487-509.
- Moreno, A., & Torres, S., (2016). *Big Data en los Sistemas de Recomendación*, Trabajo fin de Máster in Big Data & Business Intelligence, Universidad de Zaragoza, Next International Business School.
- Mata, Emili (2015). "Social management" y Big data, Harvard Deusto. https://www.harvard-deusto.com/social-management-y-big-data