# SISTEMAS DE RECOMENDACIÓN

JESÚS DANIEL HERNANDEZ LONDOÑO
502182622

UNIVERSIDAD DE PUERTO RICO
MAYAGÜEZ CAMPUS
2019

# Introducción

A medida que el tiempo avanza el aumento de información también lo hace, las bases de datos, por ejemplo, de las compañías más grandes como Facebook, Amazon y Youtube, cuentan con un sinfín de información sobre nuestras elecciones (likes, cuantas veces escucho la misma canción o cuantas veces elijo algo), es decir, con qué frecuencia doy clic a algo en particular. En la mayoría de los casos para el usuario se hace tedioso escoger entre tanta información que nos brinda internet, ¿qué quiero escuchar?, ¿qué quiero comprar?, ¿qué quiero ver?, por ende, las compañías utilizan técnicas que tratan de sugerir contenidos de acuerdo con los gustos y necesidades de los usuarios.

Netflix, por ejemplo, organizó un desafío “Netflix prize” con el objetivo de producir un sistema de recomendación que funcionará mejor que su propio algoritmo, ofreciendo como premio 1 millón de dólares, de esta manera, impulsó la investigación en esta línea, dado que 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).

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. Estos sistemas usan métodos matemáticos y estadísticos capaces de convertir los datos en conocimiento e información, surgieron a mediados de la década de los 90, dando recomendaciones a los usuarios sin personalización, es decir, ofreciendo los mismos productos y/o servicios a todos los clientes.

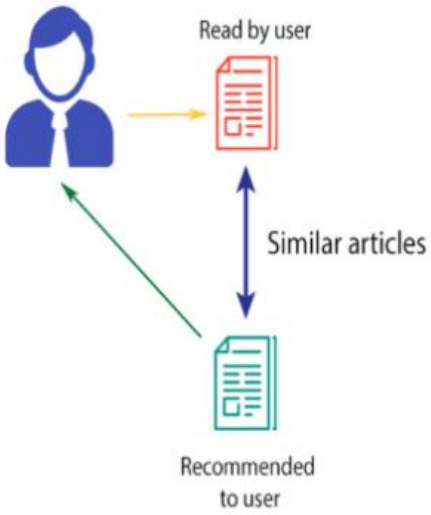
Los sistemas de recomendación se pueden clasificar, basados en contenido “Muéstrame más de lo similar a lo que me ha gustado”, filtrado colaborativos “Dime qué es popular entre mis compañeros”, basados en el conocimiento “Dime que se ajusta a mis necesidades”, e híbridos (mezcla de los anteriores).

# 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, para que estos sistemas estén enriquecidos se debería tener información adicional a la que valora el usuario, por ejemplo, nivel económico, trabajo, hobby, estado familiar, búsquedas por internet, entre otras, sin embargo este tipo de información en la práctica es difícil de obtener dado que se obtiene acceso solamente a las valoraciones echas en la compañía a la adquiere productos y/o servicios.

## Basados en contenido (content based filtering)

Este sistema analiza información previamente valorada por un usuario y construye un perfil de sus intereses, dando atributos/características con las que se describe los productos, y recomendando en base a sus gustos y no de perfiles de otros usuarios, dicho de otra forma "*muéstrame más cosas como las que me han gustado*" (Rodrigo, Joaquín. 2018).



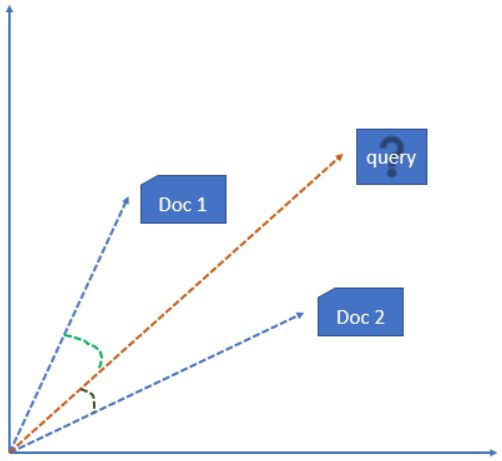

Algunas limitaciones se describen a continuación sin embargo hay entornos en donde se consigue dar buenos resultados como atributos con palabras clave, autor, temática, entre otros.

- 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.

Algunos algoritmos o técnicas usadas para este tipo de sistema son:

- Modelo de espacios vectoriales basados en palabras clave

Esta técnica representa un documento como un vector de pesos, donde cada ponderación indica el grado de asociación entre el documento y el término.




- Naïve Bayes.
- TF-IDF.

  El TF-IDF (Term Frequency - Inverse Document Frequency) o frecuencia de documento - frecuencia inversa de término, es un algoritmo que asigna la importancia a una palabra en función del número de veces que aparece en el documento, y se define como sigue:

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

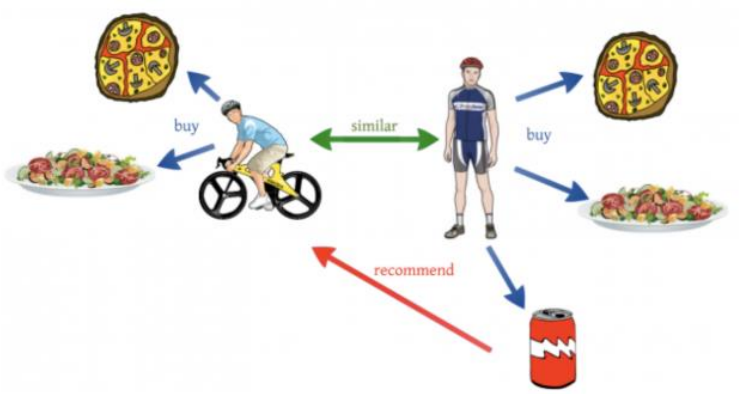
**TF-IDF**
Term $x$ within document $y$

$tf_{x,y}$ = frequency of $x$ in $y$
$df_x$ = number of documents containing $x$
$N$ = total number of documents

Donde TF(t) es la frecuencia con la que una palabra aparece en un documento, TF(t) = (Número de veces que el término t aparece en un documento) / (Número total de términos en el documento) y IDF(t) es la medida de cuán significativo es ese término, IDF(t) = log(Número total de documentos / Número de documentos con el término t).

## Filtrado colaborativo (collaborative filtering)

El Filtrado colaborativo analiza información que le ha gustado a otros usuarios con gustos similares al usuario en investigación, dicho de otra forma, "*muéstrame cosas que le hayan gustado a gente parecida a mí*" (Rodrigo, Joaquín. 2018).




Este método cuenta con algunas limitaciones:

- Un nuevo usuario se une, ya que, hasta que no se disponga de un perfil sobre sus gustos no se puede hacer recomendaciones.
- Los gustos de los usuarios pueden cambiar en el tiempo. Supongamos que hay dos usuarios, el usuario A compro muchos video juegos y un tiempo después tiene hijos, y empieza a comprar pañales y cosas para los bebes, por otro lado, hay un usuario B que disfruta de los mismos juegos que en algún momento le gustaron a A, ahora B recibe recomendaciones sobre mejores marcas de pañales.

En este método hay dos tipos de filtrado colaborativo, métodos basados en memoria y métodos basados en modelos:

- Métodos basados en Memoria:

  Este método también se conoce como algoritmos de filtrado colaborativo basados en los vecinos, en los que las calificaciones de las combinaciones de elementos de usuario se predicen en función de sus vecinos, estos vecinos se definen como:

  - Usuario - Usuario (User - User)

    Este tipo de filtrado busca perfiles similares al usuario en investigación con el objetivo de recomendar un ítem que el usuario no ha visto, utilizando las valoraciones de los otros usuarios “similares” a él, sobre el ítem en cuestión, dicho de otra forma, recomendar elementos que son más populares entre los otros usuarios “similares” al él.

  - Ítem – ítem

    Este tipo de filtrado recomienda basado en el estudio de las valoraciones de los ítems, utilizando las similaridad entre artículos en función del perfil de valoraciones que ha recibido (Rodrigo, Joaquín. 2018), puede parecer similar al basado en contenido, la diferencia es que el ítem no está definido por sus atributos (basado en contenido) sino por las valoraciones.

## Métodos basados en Modelos:

Estos métodos se utilizan en el contexto de modelos predictivos, es decir se usan modelos como árboles de decisión, modelos basados en reglas, métodos bayesianos y modelos de factores latentes, estos últimos tienen un alto nivel de cobertura. La precisión de los modelos se puede medir de muchas, sin embargo, se usará la RMSE, que es la raíz cuadrada de la media de los errores al cuadrado.

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

### Factores Latentes

Estos modelos aprovechan métodos de reducción de dimensionalidad que se usan para representar los datos subyacentes en un pequeño número de dimensiones, en otras palabras la idea básica es rotar el sistema de ejes, de modo que se eliminen las correlaciones por parejas entre dimensiones y aprovechar las redundancias de datos en datos altamente correlacionados, por ejemplo el Análisis de componentes principales (PCA) y la Descomposición de valores singulares (SVD) son un ejemplo de estos modelos.

#### Descomposición de valores singulares (SVD)

Es una forma de factorización matricial, definida como sigue

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

Donde $\mathrm{R}$ es la matriz de calificaciones y estas tres matrices $Q_{\mathrm{k}}$, $\Sigma_k$ y $\mathrm{P_k}$, son de tamaño m x k, k x k y n x k. Las matrices $Q_{\mathrm{k}}$ y $\mathrm{P_k}$, son ortogonales y contiene respectivamente los k vectores propios más grandes de $\mathrm{RR^T}$ y $\mathrm{R^TR}$. La matriz $\Sigma_k$ contiene las raíces cuadradas (no negativas) de los k valores propios más grandes de la matriz $\mathrm{R}$ a lo largo de su diagonal. Posterior a ello se redefine:

$$U = Q_k \Sigma_k$$
$$V = P_k$$
$$R = UV^T$$

El objetivo de este proceso de factorización es descubrir matrices U y V con columnas ortogonales tal que minimice

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

subject to:

Columns of $U$ are mutually orthogonal

Columns of $V$ are mutually orthogonal

En resumen:

| Approach | Conceptual Goal | Input |
|---|---|---|
| Collaborative | Give me recommendations based on a collaborative approach that leverages the ratings and actions of my peers/myself. | User ratings + community ratings |
| Content-based | Give me recommendations based on the content (attributes) I have favored in my past ratings and actions. | User ratings + item attributes |
| Knowledge-based | Give me recommendations based on my explicit specification of the kind of content (attributes) I want. | User specification + item attributes + domain knowledge |

Sacado del libro: Recommender Systems.[1].

## Medidas de Similitud

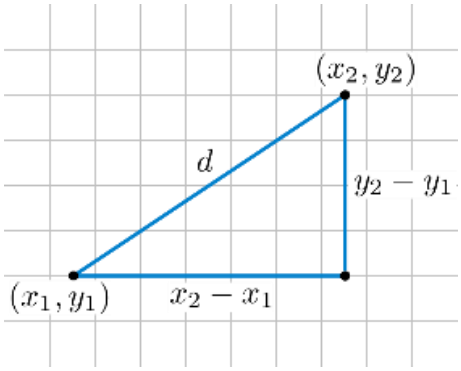
Se ha mencionado en el trascurso del documento sobre la similaridad entre dos ítems o usuarios, cuando más se asemejan dos observaciones más próximas estarán, por consiguiente, se emplea el termino distancia para determinar qué tan similares son dos observaciones, entre algunas por mencionar: la distancia euclídea, correlación, distancia coseno e índice Jaccard.

### Distancia Euclidiana

Matemáticamente la distancia euclidiana entre dos puntos es la distancia o la longitud del segmento que los une y se calcula usando el Teorema de Pitágoras.

Sean p_1 y p_2 dos puntos en el plano con coordenadas (x_1, y_1) y (x_2, y_2) respectivamente, la distancia entre p_1 y p_2 se define como:

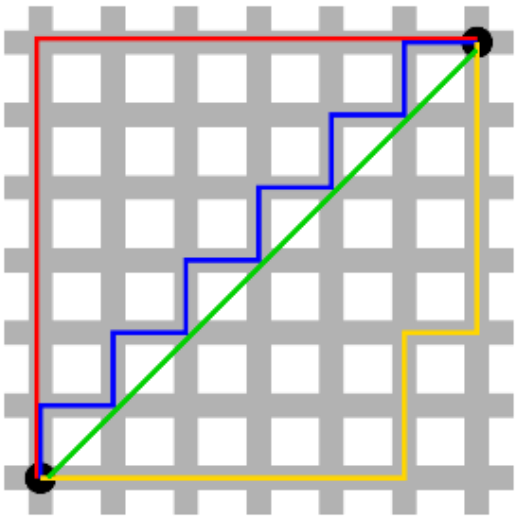
$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



## Distancia Manhattan

La Distancia Manhattan entre dos puntos considerada por Hermann Minkowski en el siglo XIX, también conocida como taxicab metric, rectilinear distance o L1 distance, es la sumatoria de las diferencias absolutas de sus coordenadas, es decir, la longitud de cualquier camino que los una mediante segmentos verticales y horizontales.

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



## Correlación

La Correlación es una medida lineal en la que se busca indicar el grado de relación entre dos variables, esta puede ser de diferentes tipos, Pearson, Spearman, Kendall, entre otras.
La correlación de Pearson funciona bien con variables cuantitativas asumiendo distribución normal, sin embargo, se menciona que sigue siendo bastante robusta si no se considera normalidad, su limitación es que es más sensible a los valores extremos que las otras dos alternativas.

La correlación de Spearman es un método no paramétrico, se emplea cuando los datos son ordinales, de intervalo, o bien cuando no se satisface la condición de normalidad.

La correlación de Kendall es otra alternativa no paramétrica para el estudio de la correlación que trabaja con rangos. Se emplea cuando se dispone de pocos datos.

### Correlación de Pearson

Es un índice definido como sigue:

$$\rho = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$ , $\text{cov}(X,Y) = E(XY) - (EX)(EY)$

bajo las siguientes condiciones y características.

-La relación que se quiere estudiar entre ambas variables es lineal (de lo contrario, el coeficiente de Pearson no la puede detectar).

-Las dos variables deben de ser cuantitativas.

-Asumir Distribución Normal.

-Toma valores entre [-1,1], siendo:
ρ=1 una correlación positiva perfecta, es decir, cuando una de ellas aumenta, la otra también lo hace en proporción constante (Relación Directa).
ρ=-1 una correlación lineal negativa perfecta, es decir, cuando una de ellas aumenta, la otra disminuye en proporción constante (Relación Inversa).
ρ=0 no existe relación lineal, sin embargo, no implica que las variables son independientes puede existir relaciones no lineales entre las variables.
$0 < \rho < 1$, existe una correlación positiva.
$-1 < \rho < 0$, existe una correlación negativa.

No varía si se aplican transformaciones a las variables.

Es sensible a outliers, por lo que se recomienda en caso de poder justificarlos, excluirlos del análisis.

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

  Esta medida se usa cuando se pretende determinar la similitud entre observaciones de tipo binaria “0”, “1”. Sean A y B dos objetos, cada uno con n atributos binarios, se define el SMC como sigue:

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

| | | A | |
|---|---|---|---|
| | | 0 | 1 |
| B | 0 | $M_{00}$ | $M_{10}$ |
| | 1 | $M_{01}$ | $M_{11}$ |

M_ {11} es el número total de atributos donde A y B tienen un valor de 1.
M_ {01} es el número total de atributos donde el atributo de A es 0 y el atributo de B es 1.
M_ {10} es el número total de atributos donde el atributo de A es 1 y el atributo de B es 0.
M_ {00} es el número total de atributos donde A y B tienen un valor de 0.

Dicho de otra forma, 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.

Ahora la distancia será 1-SMC.

## Índice Jaccard

El índice Jaccard también es una medida de similitud y se define como sigue:

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

es muy similar al Coeficiente de coincidencia simple (SMC), la diferencia radica en que no se considera M_ {00}, es decir, SMC considera como coincidencias tanto si el atributo está presente en ambos objetos A y B como si el atributo no está en ninguno de los dos, mientras que Jaccard solo cuenta como coincidencias cuando el atributo está presente en ambos.

Considere un 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.

Ahora la distancia será 1-J.

## Similitud del Coseno

Es una medida de similitud entre dos vectores distintos de cero, midiendo el coseno del ángulo entre ellos. Sean A y B dos vectores entonces su similitud será:

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

Si el ángulo es 0° es decir los dos vectores son paralelos su coseno es 1.

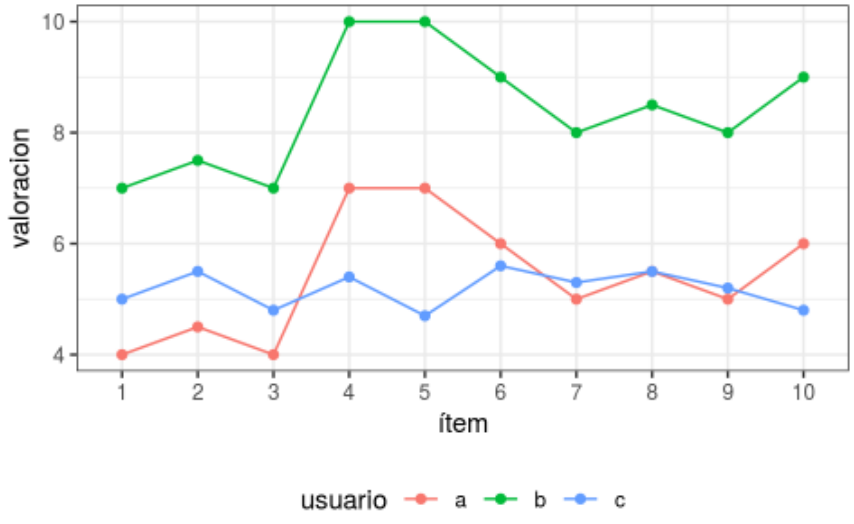
Si el ángulo es 90° es decir los dos vectores son perpendiculares su coseno es 0.

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

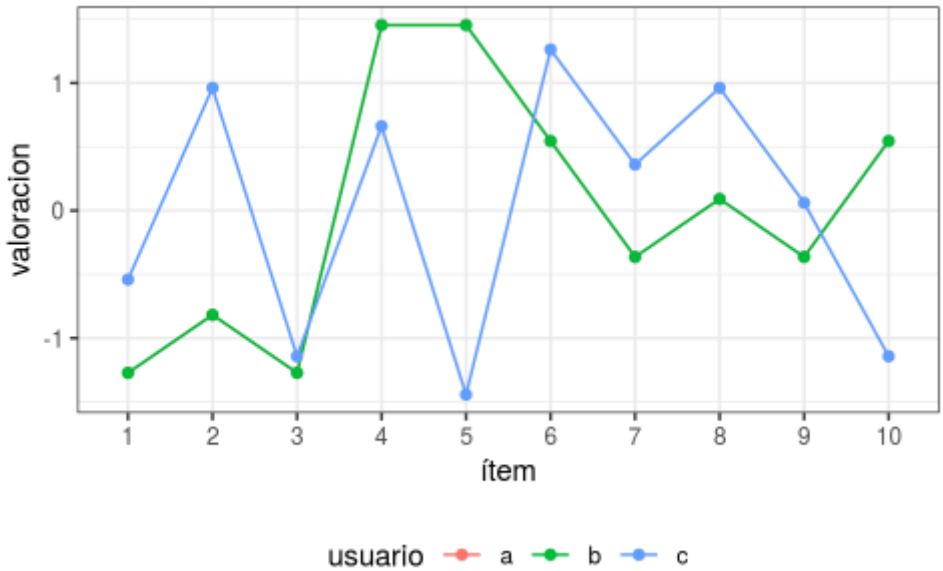
Aunque tome valores negativos la similaridad se usa particularmente en espacio positivos donde son máximamente "similares" si son paralelos y máximamente "diferentes" si son perpendiculares, además si los vectores se normalizan restándoles la media, la medida recibe el nombre de coseno centrado y es equivalente a la correlación de Pearson.

## Estandarización de valoraciones

La estandarización usualmente es usada para evitar que las variables tengan más pesos que otras, es decir, cuando aplique un algoritmo de selección de variables todas estén en las mismas condiciones de ser elegidas. Para el caso de recomendaciones, las valoraciones por el usuario, aunque son importantes el patrón o tendencia de estas también lo es.



Los usuarios a y b tienen exactamente el mismo patrón, la única diferencia es que las valoraciones del usuario b están por encima de las del usuario a 3 unidades. Usando la distancia euclidiana, los usuarios b y c son los más similares a pesar de que sus patrones son mucho más disparejos, para evitar que se oculten estos patrones se puede usar la correlación de Pearson, Kendall o estandarizar los datos.




Obteniendo que el perfil de los usuarios a y b pasan a ser idénticos, por lo que se superponen en la imagen, como se esperaba.

## Aplicación

Se anexa el documento de jupyter.

# Referencias


1- Charu C. Aggarwal, *Recommender Systems.* IBM T.J. Watson Research Center Yorktown Heights, NY, USA

2- Terveen, L., & Hill, W., (2001). Beyond Recommender Systems: Helping People Help Each Other. *HCI in the New Millennium*, 1, pp. 487-509.

3- 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.

4- Mata, Emili (2015). "Social management" y Big data, Harvard Deusto.

https://www.harvard-deusto.com/social-management-y-big-data

5- https://github.com/JoaquinAmatRodrigo/Estadistica-con-R/blob/master/PDF_format/39_Sistemas_de_recomendacion_con_R.pdf

6- https://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Pearson

7- http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/sdaugherty/similarity.html

8- https://en.wikipedia.org/wiki/Simple_matching_coefficient#cite_note-1

9- https://github.com/rounakbanik/movies/blob/master/movies_recommender.ipynb