ARTÍCULO ENVIADO

¿Cómo es que Netflix lee mi mente?

José E. Rojas-Macedo, Ángel Hernández-Castañeda, René Arnulfo García Hernández y Yulia Nikolaevna Ledeneva

Introducción

Supongamos que acabas de ver una película que te fascinó, y al día siguiente, al abrir Netflix, te encuentras con una lista de recomendaciones que parece haber sido hecha a la medida de tus gustos. ¿Brujería? No exactamente, detrás de estas recomendaciones se encuentra un algoritmo de recomendación que analiza tus preferencias y las compara con las de otros usuarios o películas para sugerirte contenido que podría interesarte.

Y no solo sucede en Netflix, dentro de la era digital, la personalización de servicios se ha convertido en una característica esencial para ofrecer una experiencia de usuario superior [Deloitte, 2024]. Desde las plataformas de streaming, hasta algunos *e-commerce*, usan estos algoritmos que nos ofrecen contenido, productos o servicios basados en nuestros gustos.

El objetivo de este artículo es desvelar de forma simple una aproximación a la construcción de estos sistemas, los tipos de datos que se emplean y los pasos que se utilizan para generar predicciones.

¿Cuál es la necesidad de estos sistemas?

A esta altura, sabrás que vivimos en una era digital en la que la cantidad de información disponible es abrumadora, y a la hora de tomar decisiones existe una sobreabundancia de opciones. Desde ese momento en el que te sientas en el sofá de tu casa y decides qué película ver, hasta el momento en el que decides qué producto comprar en línea, existe una saturación de alternativas. Con tantas opciones disponibles, uno pensaría que más es mejor, ¿cierto? Pero en realidad, sucede lo contrario: cuando uno se enfrenta a demasiadas alternativas, nuestra capacidad para tomar una decisión se ve comprometida, de acuerdo con el estudio y libro de [Schwartz et al., 2005], esto se conoce como **"parálisis por exceso de opciones"**.

De hecho existió un estudio dentro de un supermercado en el que se ofrecían 24 variedades de mermelada, y otro en el que se ofrecían solo 6. Aunque la mesa con 24 variedades atrajo a más personas, la mesa con 6 variedades fue la que generó más ventas. Volviendo a lo que se comentó antes, esto se debe a que cuando se ofrecen demasiadas opciones, las personas se sienten abrumadas y se reduce su capacidad de decisión [Iyengar and Vidal, 2011].

Justo en este punto, es donde entra a la cancha el sistema de recomendación. Plataformas como Amazon, Netflix o Spotify utilizan algoritmos inteligentes para ayudarte a reducir la cantidad de opciones y mostrar solo lo que esté dentro de tus preferencias (Figura 1). De esta forma, el beneficio de estos sistemas no solo es aumentar la facilidad con la que tomamos una decisión, sino que también mejora la experiencia en la que consumimos servicios o productos.

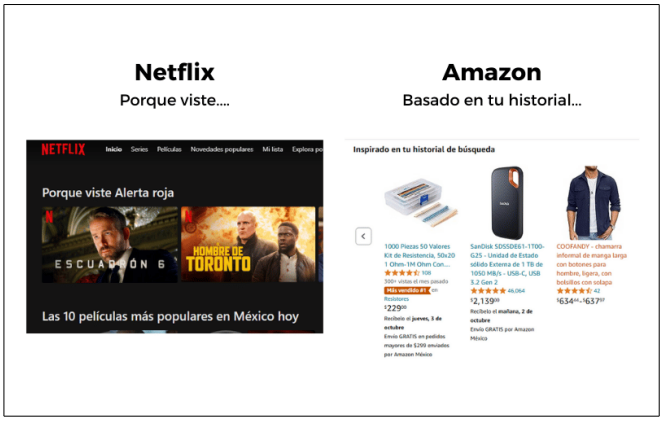


Figura 1. Utilización de algoritmos de recomendación en plataformas digitales

Sistemas de recomendación

Un sistema de recomendación, a grandes rasgos, es un algoritmo que predice la preferencia o interés de un usuario sobre un conjunto de elementos, tales como películas, productos o servicios. Estos sistemas se basan en la premisa de que si un usuario ha disfrutado de un conjunto de elementos en el pasado, es probable que disfrute de otros elementos similares en el futuro.

Un algoritmo representa una secuencia de instrucciones precisas que permiten resolver un problema o ejecutar una tarea específica, como la receta para preparar un pastel. En los sistemas de recomendación, el algoritmo recibe como entrada un conjunto de datos, en este caso, puede incluir las películas que has visto y las valoraciones que has dado sobre el contenido. Con dicha información, el algoritmo busca patrones entre tus datos y los de miles o millones más para sugerir nuevos elementos (Figura 2).

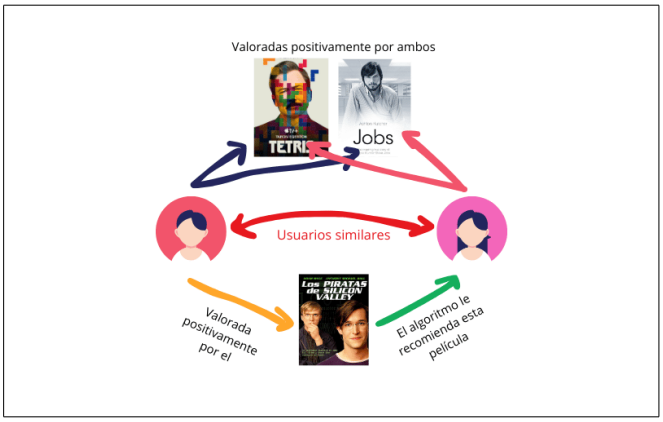


Figura 2. Funcionamiento un sistema de recomendación

En otras palabras, se basan en el comportamiento pasado para predecir lo que te gustará en el futuro. Si alguna vez has calificado una película en Netflix o repetido una canción en Spotify, estas proporcionando datos de comportamiento explícitos e implícitos al sistema respectivamente e influyendo en las recomendaciones que recibes. Existen diferentes métodos para la creación de sistemas de recomendación, sin embargo, el que se propone en este artículo es el **Filtrado Colaborativo** (FC).

En un nivel básico, el FC cuenta con dos enfoques:

1. **Basado en usuarios:** Encuentra usuarios que tengan preferencias similares a un usuario para recomendarle películas que estos hayan disfrutado.

2. **Basado en elementos:** Encuentra películas que tengan características similares a las que un usuario ha disfrutado para recomendarle otras similares.

En ambos casos, los algoritmos buscan patrones en los datos que pueden ser aprovechados para generar recomendaciones acertadas [Papadakis et al., 2022].

Para este artículo, se utilizó el enfoque basado en elementos, el cual, como se mencionó, analiza las características de las películas que un usuario ha disfrutado, para que el algoritmo busque películas que hayan sido valoradas de manera similar y que compartan características, al encontrar estos patrones, el sistema devuelve las opciones encontradas.

Los sistemas de recomendación analizan miles o millones de datos para detectar patrones y sugerirte sugerencias personalizadas, o lo que también se conoce como hiperpersonalización.

Metodología

Si actualmente cursas la preparatoria o tienes más de 20 años, muy seguramente te haya tocado conocer la cadena de Blockbuster, una tienda de alquiler de películas que se hizo famosa en los años 90 y principios de los 2000. Todos los días cientos de clientes entraban, miraban a los estantes y elegían una película para llevársela a su casa, incluso algunos empleados te recomendaban películas que podrían interesarte con base a tus preferencias que tú les contaras (justo lo que hace un algoritmo de FC). Era un servicio que solo algunos daban porque imagínate que todos tuvieran esa habilidad manual para recomendar películas a todos los clientes, además de la complejidad, sería un caos. La desaparición de esta cadena se debió a la llegada de Netflix, una plataforma que supo aprovechar este problema para ofrecer un servicio de recomendación personalizado, además a otros factores de su modelo de negocio desactualizado [Davis and Higgins, 2013].

En esta sección, se va a describir la metodología utilizada para la construcción de un sistema de recomendación basado en filtrado colaborativo. El proceso se divide en las siguientes etapas:

1. **Recolección y preparación de datos:** El primer paso consistió en conseguir un conjunto de datos de películas que incluyera información sobre los usuarios, películas y valoraciones de los usuarios por película, para ello, se utilizó el conjunto de datos de MovieLens

2. **Análisis exploratorio:** Antes de intentar hacer recomendaciones, primero se deben entender los datos, es decir, revisar todas las opiniones que existen y la cantidad que hay en el conjunto.

3. **Organización de la información:** Después de que se hace un análisis de los datos, se procede a su organización, en donde se combinaron los datos de las películas y las valoraciones para formar una tabla gigante, algo similar a lo que puedes ver en Excel, con filas y columnas. Pero existe un detalle, no todos los usuarios han visto todas las películas, este es el reto.

4. **Desarrollo del modelo:** El desafío consiste en rellenar esos huecos, para ello, el sistema debe literalmente adivinar qué tanto te gustaría *Tetris* si no la has visto, basado en lo que ya has observado tú y otros usuarios. Para ello, se buscan patrones en las características de las películas que has visto

y se comparan con las de otros usuarios para hacer recomendaciones, de esta forma, si viste *Steve Jobs* y *The Social Network*, el sistema podría recomendarte *Tetris*, porque hablan sobre la vida de un emprendedor y la tecnología.

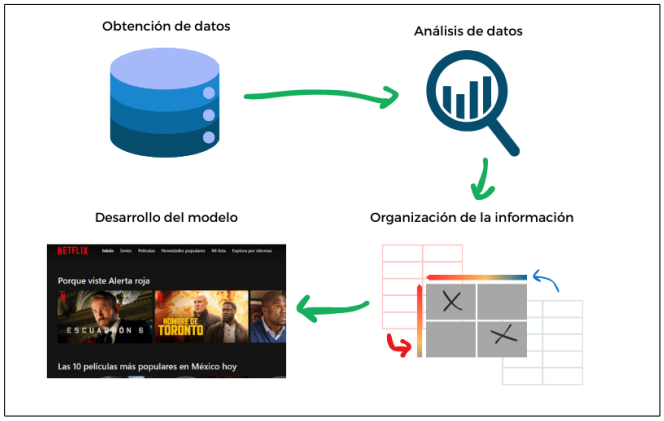


Figura 4. Ilustración gráfica de la metodología utilizada

Recolección y preparación de datos

Como se había hablado, el primer paso es encontrar un conjunto de datos correcto que contenga la información relevante para realizar este tipo de sistemas. En el caso de este experimento, se utilizó un conjunto de datos de una plataforma de recomendaciones llamada MovieLens (ver Figura 3), que fue desarrollada por la Universidad de Michigan en 1998, y que ha sido utilizada en la investigación de sistemas de recomendación [Harper and Konstan, 2015].

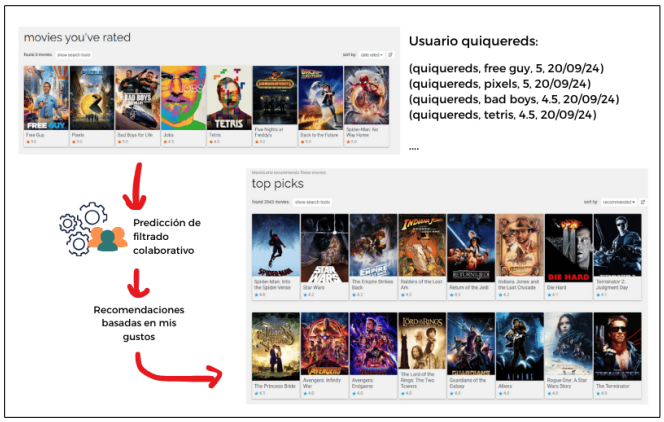


Figura 3. Plataforma de MovieLens utilizando un sistema de recomendación

El conjunto que se utilizó representa una versión ligera de toda la base de datos, se incluyen alrededor de 100,000 valoraciones de 600 usuarios aplicadas a 9,000 películas hasta 2018. Imagina que es como un cuaderno donde tienes apuntado lo han han visto tú y tus amigos, así como la calificación que le dan a la película. Toda esta información se almacena en dos archivos clave de este conjunto de datos, que contienen:

- **Listado de calificaciones (ratings.csv):** ¿Que películas vieron los usuarios y qué calificación les dieron?
- **Listado de películas (movies.csv):** Detalles de la película como su nombre, géneros, etc.

Análisis de datos

El siguiente paso es conocer el terreno en el que se está trabajando, y para ello, se realizó un análisis exploratorio de los datos recopilados.

¿A que se refiere esto? Volvamos al ejemplo del cuaderno de notas, ahora el siguiente paso es revisarlo. Para este punto, debemos entender cuántas películas han sido calificadas por cada usuario, cuáles son las más vistas y que tipo de valoraciones son las más comunes. Todo este proceso, ayudará a identificar patrones importantes en los datos.

Durante el primer análisis exploratorio, se encontró la siguiente información:

- El conjunto de datos contiene 100,836 valoraciones de 610 usuarios.
- Existen 9,724 películas en el dataset.
- El promedio de valoraciones por usuario es de 165 valoraciones.
- El promedio de valoraciones por película es de 10 valoraciones.

De manera gráfica, en la Figura 5, se muestra la distribución de las valoraciones, lo cual es importante para determinar si hay problemas de tendencia en los datos. En este caso, se observa que la mayoría de las valoraciones se encuentran en el rango de 3 a 4 estrellas, lo cual es un buen indicador de que el conjunto está equilibrado.

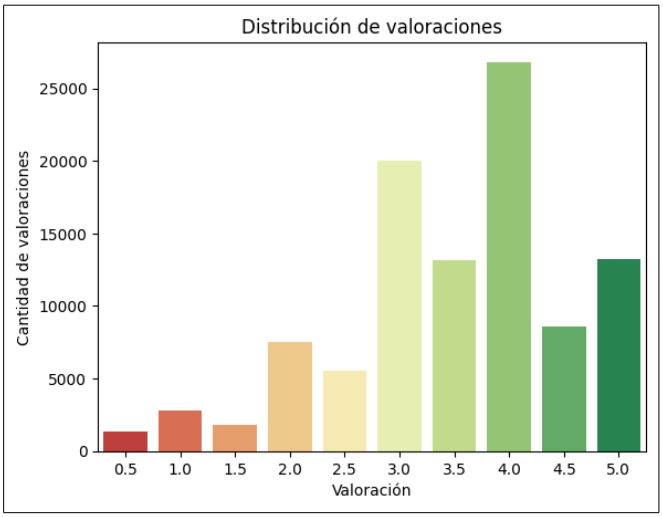


Figura 5. Distribución de las valoraciones en el conjunto de datos

Adicionalmente, será bastante común encontrar que aquellas películas que son altamente populares cuenten con muchas valoraciones, mientras que aquellas que no lo son, cuenten con pocas. Comprender esto es muy útil ya que ayuda a determinar si existe información suficiente para dar recomendaciones precisas o si hay escasez de información (lo que provocaría una falsa representación).

En el análisis, se encontró que la película con mejor calificación es *Lamerica (1994)*, mientras que la peor calificación la tiene *Gypsy (1962)*, profundizando un poco más, se encontró un detalle, la película de *Lamerica* solo tiene 2 valoraciones, entonces hay una falsa representación. Si una película tiene una calificación muy alta, pero que solo ha sido evaluada por una minoría de la muestra de usuarios, no se puede considerar como una buena recomendación.

Para mitigar este problema, se utilizó el promedio bayesiano de las valoraciones, en donde en lugar de confiar únicamente en las calificaciones de los usuarios a una película, el promedio bayesiano combina estas calificaciones con el promedio global de todas las películas. Esto se hace para evitar que películas con muy pocas valoraciones tengan un impacto desproporcionado. Con este ajuste:

- La película con mejor calificación es *Shawshank Redemption (1994)*.
- La película con peor calificación es *Speed 2: Cruise Control (1997)*.

El siguiente paso fue el análisis de los géneros de las películas, con el objetivo de determinar qué géneros son los más populares en el conjunto, de esta forma en la Figura 6 se muestra la distribución, donde se observa que los géneros más populares son *Drama*, *Comedia*, *Thriller* y *Acción*.

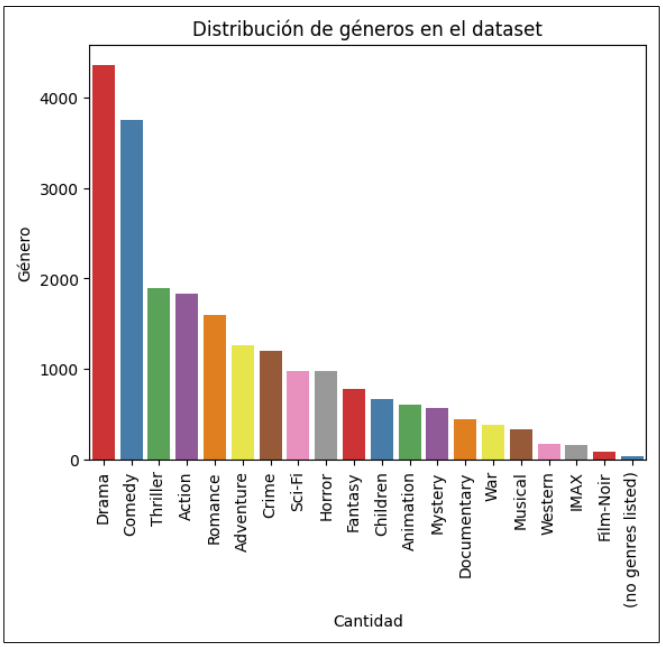


Figura 6. Distribución de los géneros en el conjunto de datos

Organización de la información

Hasta el momento, se han recolectado y analizado los datos, y tenemos dos listas, una con las películas y otra con las valoraciones de los usuarios, lo cual, en un sistema de recomendación no es útil, ya que necesitamos una tabla que contenga toda la información en un solo lugar, a lo que denominaremos *matriz usuario-película*.

Esta matriz se puede entender como una tabla de Excel enorme, donde cada fila representa a un usuario, y cada columna, a una película. Dentro de la tabla, las celdas contienen las valoraciones que los usuarios han dado a las películas, y los espacios vacíos representan las películas que los usuarios no han visto.

Es decir, si Juan ha visto *The Social Network* y le ha dado 5 estrellas, en la tabla, en la fila de Juan y la columna de *The Social Network*, se colocará un 5. Si Juan no ha visto *Tetris*, la celda correspondiente estará vacía. El objetivo del FC basado en elementos es determinar qué película le gustaría a Juan, basado en las películas que ha visto y las valoraciones que ha dado.

Para fines de eficiencia, se implementó un "mapa"de películas y usuarios en el que se asigna un identificador único a cada película y usuario, de esta forma, en lugar de trabajar con nombres, se trabaja con números, lo que permite al algoritmo determinar rápidamente dónde están ubicados los datos en la matriz.

Desarrollo del modelo

Finalmente, se llega a la parte más interesante, el desarrollo del modelo de recomendación. En este caso,

se utilizó una técnica llamada **K-Nearest Neighbors** (KNN).

Al enfoque de recomendación de películas basado en otra película, le llamaremos **recomendaciones item-item**, lo que consiste en la identificación de las características de las películas y las relaciones que exista entre todo el conjunto. El sistema busca películas que hayan sido valoradas de forma similar por muchos usuarios, encuentra similitudes, realiza comparación de características y devuelve las recomendaciones.

El algoritmo KNN se puede comprender como una red de tus amistades, en este caso, entre películas: aquellas que han tenido valoraciones similares por parte de muchos usuarios, se consideran amigas. Si te gustó *El señor de los anillos*, el sistema podría recomendarte *Harry Potter y La piedra filosofal* o *El Hobbit*, porque comparte características con la primera, están en la misma categoría de fantasía (red de amigas) y han sido valoradas de forma similar.

Lo anterior, fue lo que se obtuvo exitosamente como resultado del experimento, se tomó la película *Avengers: Era de Ultron* como película base y el algoritmo recomendó *Guardianes de la Galaxia*, *Thor: Un Mundo Oscuro* y *Capitán America: Soldado del Invierno*, ¿Por qué? Volvemos al párrafo anterior, comparten características de que tratan de superhéroes, son del mismo universo cinematográfico (Marvel) y han sido valoradas de forma similar por los usuarios.

Conclusiones

Los sistemas de recomendación, como los que incorpora Netflix, en resumen son un ejemplo de cómo la tecnología puede ser utilizada para mejorar la experiencia del usuario utilizando los datos que generan. En este artículo, se ha presentado una aproximación a la construcción de un sistema de recomendación basado en filtrado colaborativo, que aunque su proceso es complejo, el objetivo es simple: hacer que disfrutes de tu contenido favorito, y aunque este algoritmo está lejos de ser incluso parecido al que utiliza Netflix, es un buen punto de partida para entender cómo funcionan estos sistemas y responder a tu pregunta de ¿cómo es que Netflix lee mi mente?

Referencias

[Davis and Higgins, 2013] Davis, T. and Higgins, J. (2013). A blockbuster failure: how an outdated business model destroyed a giant.

[Deloitte, 2024] Deloitte (2024). Connecting with meaning - hyper-personalizing the customer experience using data, analytics, and ai.

[Harper and Konstan, 2015] Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. ACM Trans. Interact. Intell. Syst., 5(4).

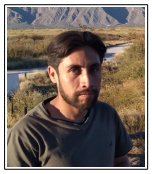
[Iyengar and Vidal, 2011] Iyengar, S. and Vidal, M. (2011). El arte de elegir: Decisiones cotidianas. Qué dicen de nosotros y cómo podemos mejorarlas. Sin colección. Grupo Planeta.

[Papadakis et al., 2022] Papadakis, H., Papagrigoriou, A., Panagiotakis, C., Kosmas, E., and Fragopoulou, P. (2022). Collaborative filtering recommender systems taxonomy. Knowledge and Information Systems, 64(1):35–74.

[Schwartz et al., 2005] Schwartz, B., Bustelo, G., and Carretero, T. (2005). Por qué más es menos: la tiranía de la abundancia. Pensamiento (Taurus). Taurus.

SOBRE LOS AUTORES

José Enrique Rojas Macedo recibió el título de Ingeniero de Software por la Universidad del Valle de México (UVM) en 2023, especializado en el desarrollo de aplicaciones móviles con Flutter, Google Cloud y Firebase. Motivado por empujar constantemente sus límites, contribución a causas sociales, compartir conocimiento y actualmente estudiando la Maestría en Ciencias de la Computación en la Universidad Autónoma del Estado de México. Sus intereses en investigación son el análisis de sentimientos, reconocimiento de patrones y minería de datos.



Ángel Hernández Castañeda recibió su Maestría y Doctorado en Ciencias de la Computación, con honores, por el Centro de Investigación en Computación (CIC) del Instituto Politécnico Nacional (IPN), en 2013 y 2017, respectivamente. Asimismo, es autor de diversas publicaciones en revistas internacionales de alto impacto. Actualmente es Profesor Investigador de Tiempo Completo en la Universidad Autónoma del Estado de México y miembro del Sistema Nacional de Investigadores (SNI) de México. Sus intereses en investigación incluyen el procesamiento de lenguaje natural, la minería de datos, el aprendizaje automático y el reconocimiento de patrones.