



Análisis de grandes volúmenes de datos

Dr. Néstor Velasco Bermeo

Tarea 1. Sistema de Recomendación

Abril 28 2024

EQUIPO 8

Joel Orlando Hernández Ramos
A00759664

María de los Ángeles Rabelero Campos
A01793031

Daniel Gómez Álvarez
A01794498

Introducción

Los sistemas de recomendación (SR) son herramientas avanzadas de inteligencia artificial y aprendizaje automático que se integran profundamente en nuestras experiencias digitales diarias. Estos sistemas están diseñados para analizar nuestras preferencias y comportamientos en línea, ofreciéndonos sugerencias personalizadas de productos, servicios y contenidos que probablemente sean de nuestro interés. En una era donde existe un sin fin de información disponible en Internet, la importancia de los sistemas de recomendación es innegable, ya que facilitan decisiones rápidas y eficientes frente a la abundancia de opciones disponibles en diversas plataformas, desde el comercio electrónico hasta los medios de comunicación digital y redes sociales.

A grandes rasgos, los SR funcionan como herramientas de filtrado de información, proporcionando a los usuarios contenido adecuado y personalizado que reduce significativamente el esfuerzo y el tiempo necesario para buscar información relevante en internet. Estos sistemas manejan dos entidades principales: usuarios y artículos, donde cada usuario proporciona una calificación (o valor de preferencia) a un artículo. Las calificaciones de los usuarios generalmente se recopilan mediante métodos implícitos o explícitos. Las calificaciones implícitas se recogen indirectamente a través de la interacción del usuario con los artículos, mientras que las explícitas se otorgan directamente por el usuario eligiendo un valor en una escala de puntos finita. Por ejemplo, un sitio web puede obtener calificaciones implícitas de diferentes artículos basadas en datos de secuencia de clics o en la cantidad de tiempo que un usuario pasa en una página web, entre otros. La mayoría de los sistemas de recomendación recopilan calificaciones de los usuarios a través de ambos métodos. Estos comentarios o calificaciones proporcionados por los usuarios se organizan en una matriz usuario-artículo llamada matriz de utilidad, que a menudo contiene muchos valores faltantes. El problema principal de los sistemas de recomendación se centra en encontrar los valores que faltan en esta matriz de utilidad, una tarea que suele ser difícil debido a la escasez inicial de la matriz, ya que los usuarios generalmente tienden a calificar solo un pequeño número de artículos.

En este sentido, la relación entre los sistemas de recomendación y los grandes datos es fundamentalmente simbiótica. Estos sistemas utilizan complejos algoritmos para predecir las necesidades y deseos de los usuarios, mejorando así continuamente la precisión de sus recomendaciones a través del análisis de grandes volúmenes de datos. Esta cantidad masiva de datos alimentan los sistemas de recomendación, proporcionando los conjuntos de datos vastos y variados requeridos para que los algoritmos sofisticados predigan e influyan en el comportamiento del consumidor de manera efectiva.

Es importante mencionar que este proceso no solo enriquece la experiencia del usuario al reducir la sobrecarga de información, sino que también optimiza indicadores relevantes comerciales como las ventas, la participación de los usuarios y las tasas de retención de clientes. La relevancia empresarial de los sistemas de recomendación es sustancial y multifacética. Al personalizar las interacciones de los usuarios, compañías como Netflix y

Amazon han demostrado que los motores de recomendación efectivos pueden llevar a aumentos dramáticos en el compromiso del usuario y los retornos financieros. Por ejemplo, las recomendaciones representan una porción significativa de la actividad del usuario en estas plataformas, correlacionándose directamente con un aumento de ingresos y una reducción en la deserción. La implementación estratégica de sistemas de recomendación puede llevar a resultados empresariales mejorados, desde tasas de clics más altas y ventas, hasta una mayor lealtad del cliente y diferenciación competitiva.

En este ensayo se profundizará en los tipos de sistemas de recomendación y evaluará el impacto de ellos las prácticas empresariales de principalmente dos industrias: e-Commerce y sistemas de Streaming. Esto con el objetivo de aprender a identificar los retos que enfrenta una empresa al tratar de utilizar grandes volúmenes de datos.

Tipos de Sistemas de Recomendación

Los sistemas de recomendación se clasifican comúnmente en varias categorías según los métodos empleados para generar las recomendaciones. Estos métodos pueden variar significativamente en su enfoque y complejidad. Cada uno cuenta con sus ventajas y limitaciones. A continuación, se describen detalladamente los principales tipos de sistemas de recomendación identificados en la literatura mostrados en la Figura 1.

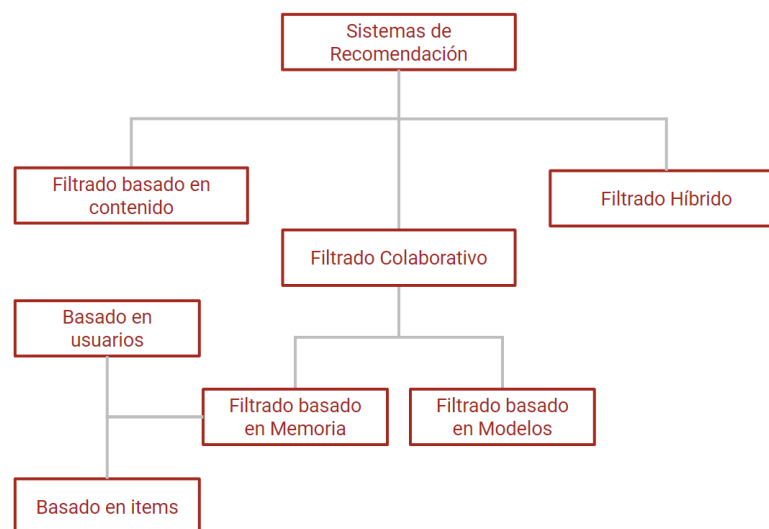


Figura 1. Esquema tipos de sistemas de Recomendación.

Filtrado basado en contenido

El filtrado basado en contenido implica recomendar artículos similares a los que un usuario ha preferido en el pasado. Este tipo de sistema analiza las características de los artículos que el

usuario ha evaluado favorablemente para construir un perfil de intereses del usuario. Por ejemplo, en el caso de los libros, estas características pueden incluir el autor y la editorial; para las películas, podrían ser el director y los actores. El sistema acumula estos datos en perfiles de ítems que, cuando un usuario muestra interés por un ítem, se utilizan para formar un perfil de usuario. Este perfil es luego empleado para recomendar nuevos ítems que coincidan con las características almacenadas. A pesar de su capacidad para adaptarse rápidamente a los cambios en las preferencias del usuario, este enfoque tiene la desventaja de requerir un conocimiento profundo de las características de los ítems, que no siempre está disponible. Además, tiene una capacidad limitada para expandir las preferencias existentes del usuario más allá de lo que ya conocen.

Filtrado colaborativo

El filtrado colaborativo se basa en la premisa de que personas con gustos y preferencias similares probablemente estarán interesadas en las mismas cosas. Este método no requiere conocimiento sobre las características de los ítems. En cambio, hace uso de la similitud entre los usuarios para hacer recomendaciones. Se puede dividir en dos subtipos principales:

Colaborativo basado en memoria

Este enfoque utiliza directamente la matriz de utilidad, que almacena las calificaciones de los usuarios a los ítems, para predecir qué tan relevante podría ser un ítem para un usuario basado en las calificaciones de otros usuarios con gustos similares. Se subdivide a su vez en:

- **Colaborativo basado en usuarios:** Se predice la calificación de un usuario para un ítem no valorado basándose en las calificaciones de otros usuarios similares que sí lo han valorado.
- **Colaborativo basado en ítems:** Se construye un "vecindario" de ítems similares a los que un usuario ha valorado positivamente, y se utilizan estas relaciones para predecir la valoración de nuevos ítems.

Colaborativo basado en modelos

Este enfoque construye un modelo matemático o estadístico (a menudo utilizando técnicas de aprendizaje automático) que predice las calificaciones basadas en la matriz de utilidad. Este modelo puede ser más eficiente en términos computacionales y manejar mejor ciertos problemas como la escasez de datos y la escalabilidad.

Sistemas de recomendación híbridos

Los sistemas híbridos combinan dos o más técnicas de los tipos anteriores para mejorar el rendimiento y compensar las limitaciones individuales de cada enfoque. Por ejemplo, un sistema híbrido puede integrar el filtrado colaborativo y basado en contenido para proporcionar recomendaciones que no solo se basen en la similitud del usuario o del contenido, sino en una combinación de ambos. Otras formas de hibridación incluyen la combinación de características, la hibridación en cascada, y la hibridación ponderada, entre otras. Este tipo de sistema es

particularmente útil para abordar el problema de arranque en frío, donde los nuevos usuarios o ítems no tienen suficientes datos para hacer recomendaciones efectivas. La Tabla 1 muestra los diferentes tipos de filtraciones híbridas, así como una pequeña descripción.

Hybrid methods	Description
Meta-level	A pre-learned model is used as an input to another recommender system
Feature combination	The features of one recommender system are injected into another
Feature augmentation	The result of one model is applied as an input to another
Mixed hybridization	The output of different recommender systems are mixed, and the combined result is given as a recommendation
Cascade hybridization	One system improvises the output of another
Switching hybridization	Select one recommender model based on the current requirement
Weighted hybridization	Ratings of different techniques are aggregated to compute a single recommendation

Tabla 1. Métodos Híbridos de Filtrado

Objetivos y Metas en las Industrias

e-Commerce

En la industria del comercio electrónico, los sistemas de recomendación tienen el objetivo primordial de mejorar la experiencia de compra al personalizar las sugerencias de productos para cada usuario. El objetivo es aumentar la tasa de conversión y el valor promedio del carrito de compras al presentar productos que no solo coinciden con los intereses del usuario sino que también complementen compras anteriores. Además, estos sistemas buscan optimizar la gestión del inventario al promover productos con mayor margen de beneficio o aquellos que necesitan ser liquidados. Asimismo, al analizar grandes volúmenes de datos sobre el comportamiento de compra y las preferencias de los usuarios, los sistemas de recomendación en e-commerce facilitan la identificación de tendencias de mercado y la adaptación rápida a las demandas cambiantes, lo que resulta en una mejor satisfacción del cliente y fidelidad a largo plazo.

Las empresas de ventas electrónicas utilizan primordialmente sus propios SR, desarrollados por ellos. Debido a que tienen suficientes datos de sus clientes, lo que les facilita el acceso y permite una recomendación inteligente y precisa. Pero empresas no especializadas en el comercio electrónico han adoptado software no creado por ellos, plataformas de recomendación como servicio. Revisaremos algunas plataformas populares.

Algolia Recommend

Algolia es una plataforma de búsqueda principalmente, pero también ofrece un SR. Las recomendaciones se basan en modelos supervisados de aprendizaje automático que se entrenan a partir de los datos de sus productos y las interacciones de los usuarios.

Recommend usa dos tipos de algoritmos diferentes: el filtrado colaborativo y el filtrado basado en el contenido.

El filtrado colaborativo analiza los eventos de usuario de los últimos 30-90 días. Recommend crea una tabla de userToken y objectID que muestra cuántas veces interactúa cada usuario con cada registro. Recommend utiliza un algoritmo de filtrado colaborativo para encontrar otros registros que sean similares o que se compren juntos con frecuencia:

- Similares si el mismo conjunto de usuarios interactúa con ellos.
- Frecuentemente comprados juntos si el mismo conjunto de usuarios los compró.

El filtrado basado en el contenido analiza atributos clave de los objetos, como sus títulos o descripciones, para encontrar similitudes.

Algolia Recommend construye modelos a partir de los eventos de los usuarios, basándose en algoritmos de aprendizaje automático supervisado. Dado un artículo visto, el modelo entrenado recomienda otros artículos relacionados.

El modelo Frecuentemente comprado juntos tiene 2 variantes, Relajado y Estricto, ambas recomiendan artículos que van bien juntos:

- Relajado. Artículos que podrían ir bien juntos.
- Estricto. Artículos que se han comprado juntos en el pasado.

Los modelos Productos relacionados y Contenidos relacionados recomiendan artículos relacionados entre sí, basándose en:

- Interacciones del usuario (eventos de clic y conversión) (filtrado colaborativo).
- Atributos (filtrado basado en el contenido). Con el filtrado basado en el contenido, puede mostrar contenido relacionado en lugar de productos relacionados.

El filtrado basado en el contenido crea recomendaciones aún más relevantes que confiar únicamente en el filtrado colaborativo y le permite mostrar recomendaciones cuando no tiene suficientes eventos de clics y conversiones. Con el filtrado basado en el contenido, puede maximizar la exposición de su catálogo para que los usuarios reciban recomendaciones relevantes, incluso sobre artículos que no son populares.

Recombee

Es un servicio para la recomendación de productos, contenido y búsqueda inteligente. Personas y empresas con una tienda electrónica pueden usar este servicio para personalización en tiempo real para mejorar la experiencia del usuario mediante su sistema de recomendación basado en contenido, filtrado colaborativo y el comportamiento de los usuarios.

Recombee utiliza AutoML AI para optimizar los conjuntos de algoritmos. También implementa redes neuronales profundas para ayuda a predecir la próxima acción basándose en el comportamiento histórico y las reglas de negocio mediante “boosters” o filtros,

Desarrollo de una lógica de recomendación personalizada procesando el lenguaje natural de los post de producto, escanea el texto en las imágenes de los productos (OCR), y también procesa las imágenes estáticas y de hace muestreado de los videos de los productos y las reseñas

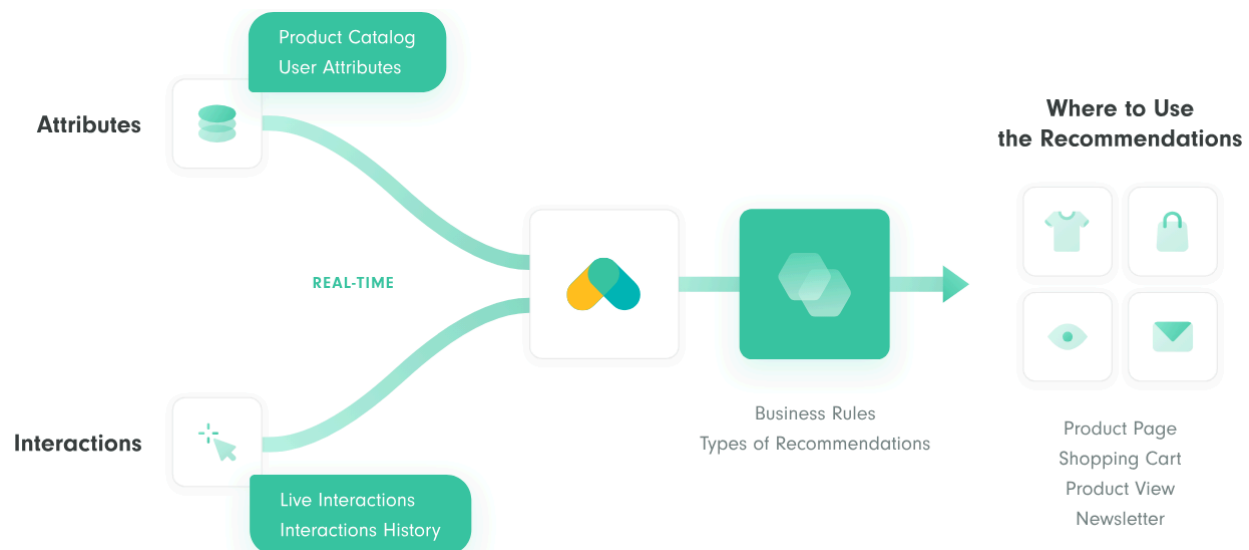


Figura 2. Arquitectura de Recomendación de Recomendaciones.

Streaming Services

Netflix

Para los servicios de streaming, los sistemas de recomendación se diseñan con el fin de maximizar la retención de usuarios y aumentar el tiempo de visualización mediante la personalización de los contenidos ofrecidos. Estos sistemas buscan entender profundamente las preferencias individuales y los patrones de consumo para sugerir películas, series o programas que mantengan al usuario enganchado y satisfecho. El objetivo es también descubrir y recomendar contenido que, aunque esté fuera de las preferencias habituales del usuario, pueda resultar de interés, fomentando así una exploración más amplia dentro del

catálogo disponible. Esto no solo mejora la experiencia del usuario, sino que también contribuye al descubrimiento de contenidos menos conocidos, aumentando su visibilidad y consumo. Además, un sistema de recomendación eficaz puede ayudar a predecir la popularidad de contenidos futuros y guiar decisiones estratégicas sobre producciones originales o adquisiciones de nuevos contenidos.

El algoritmo de recomendación de Netflix ha sido bastante popular entre las personas que estudian el análisis de datos. Netflix tiene una enorme colección de datos de usuarios y sigue recopilando más con cada nuevo usuario y actividad de los usuarios.

Tiene en cuenta numerosas métricas de este conjunto de datos para construir el algoritmo, algunas de las cuales son:

- **Historial de visualización y calificaciones.** Netflix puede hacer excelentes sugerencias según el tipo de películas que se mira y cómo se califica las películas que los usuarios han visto. Netflix se centra en el historial de visualización de un usuario con otros factores y encuentra un patrón sobre la elección de cada usuario.
- **Usuarios con gustos y elecciones similares.** Netflix usa un algoritmo de filtrado colaborativo (CF). Se basa en la idea de que si dos usuarios con un historial de calificación similar se comportarán de manera similar en el futuro. Por ejemplo, si uno de los dos usuarios similares califican una película con una puntuación alta, es probable que al otro usuario también le guste y, por tanto, Netflix se la recomendará al otro.
- **Títulos, géneros, actores, año de lanzamiento, etc.** Esto configura el algoritmo de filtrado basado en contenido (CB). Este algoritmo recomienda películas basadas en las que a los usuarios les han gustado antes. La principal diferencia con CF es que CB ofrece recomendaciones basadas no solo en la similitud por calificación, sino que se centra más en la información de los productos. Por ejemplo, recomendar películas del mismo actor que te gustaban antes. Estos factores con la elaboración de perfiles de usuario ayudan a conocer las preferencias del usuario.
- **El perfil de usuario** incluye información como la hora del día en que el usuario ve algo, el dispositivo en el que se ve, la duración que se ve, etc. Por ejemplo, si muestra un patrón de no ver por mucho tiempo, Netflix recomendará películas cortas o programas con tiempos de ejecución de episodios pequeños.

Netflix considera varios factores para elegir la imagen de un programa. También tendrá en cuenta algo básico como la cantidad de clics que recibe una determinada imagen. La imagen con más clics o popularidad se convierte en uno de los factores del algoritmo. Según las preferencias y el historial de visualización, si Netflix predice que a un usuario le gustan las comedias, mostrará la imagen de Robin Williams para la película de “Good Will Hunting”. Si al usuario le gustan las películas románticas, Netflix mostrará la imagen de Matt Damon y Minnie Driver, por los elementos de romance.

En Netflix, se usa una combinación de algoritmos con conexión (on-line) y sin conexión (offline), o una combinación de ambos donde la mayor parte de las recomendaciones se calculan fuera

de línea y se agrega algo de frescura al post-procesar las listas con algoritmos en línea que usan señales en tiempo real.

Se almacenan los resultados intermedios y sin conexión en varios repositorios para consumirlos posteriormente en el momento de la solicitud: los principales almacenes de datos Netflix utiliza son Cassandra, EVCache y MySQL. Cada solución tiene ventajas y desventajas sobre las demás.

MySQL permite el almacenamiento de datos relacionales estructurados que podrían ser necesarios para algún proceso futuro mediante consultas de propósito general. Sin embargo, la generalidad tiene el costo de problemas de escalabilidad en entornos distribuidos. Cassandra y EVCache ofrecen las ventajas de las tiendas de valor clave.

Cassandra es una solución estándar y conocida cuando se necesita un almacén sin SQL distribuido y escalable. Cassandra funciona bien en algunas situaciones; sin embargo, en los casos en los que necesitamos operaciones de escritura intensivas y constantes, encontramos que EVCache se adapta mejor.

Sin embargo, la cuestión clave no es tanto dónde almacenarlos, sino cómo manejar los requisitos, de manera que objetivos conflictivos como la complejidad de las consultas, la latencia de lectura/escritura y la coherencia transaccional se encuentren en un punto óptimo para cada caso de uso. En referencias

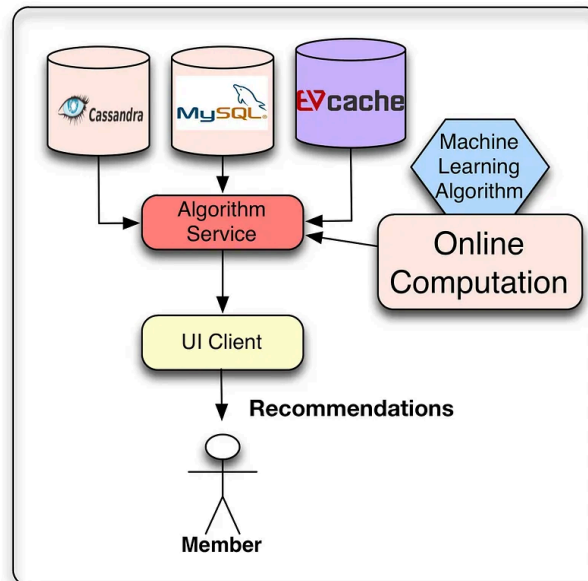


Figura 3. Arquitectura de Netflix

Conclusión

En resumen, los sistemas de recomendación no son simplemente facilitadores tecnológicos, sino catalizadores esenciales para la transformación digital y la competitividad empresarial. Su capacidad para analizar y sintetizar grandes volúmenes de datos en recomendaciones personalizadas ha demostrado ser un diferenciador crítico en el panorama comercial actual. Empresas líderes como Netflix y Amazon han probado que un enfoque sofisticado en la personalización no solo eleva la experiencia del usuario, sino que también impulsa la retención de clientes y el crecimiento de ingresos. Sin embargo, el desarrollo continuo y la optimización de estos sistemas son cruciales para mantener la relevancia en un mercado en constante evolución. Un tema que no se aborda en este ensayo pero es un desafío futuro para los desarrolladores y empresas no es solo implementar tecnologías de recomendación avanzadas, sino también asegurar que estas sean éticamente responsables y transparentes para los usuarios, garantizando así un equilibrio entre la personalización y privacidad.

Referencias

Jannach, D., & Jugovac, M. (2019). Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)*, 10(4), 1-23.

Manouselis, N., Drachsler, H., Verbert, K., & Duval, E. (2012). *Recommender systems for learning*. Springer Science & Business Media.

Roy, C., Rautaray, S.S. (2021). Challenges and Issues of Recommender System for Big Data Applications. In: Rautaray, S.S., Pemmaraju, P., Mohanty, H. (eds) *Trends of Data Science and Applications*. Studies in Computational Intelligence, vol 954 . Springer, Singapore.
https://doi.org/10.1007/978-981-33-6815-6_16

Roy, D., Dutta, M. A systematic review and research perspective on recommender systems. *J Big Data* 9, 59 (2022). <https://doi.org/10.1186/s40537-022-00592-5>

Amin B., Shahpar Y., Salman M., Seyed G., Srinivasa R., and Mohammad A.. Towards cognitive recommender systems. *Algorithms*, 13(8):176, 2020b.

frantatoman.com. (n.d.). Product recommendation engine tailored by data scientists. Recombee. <https://www.recombee.com/product-recommendations.html>

Julian McAuley, Ruining He, Chen Fang, Zhaowen Wang. Vista: A visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems, RecSys '16*, pages 309–316, New York, NY, USA, 2016. ACM.

Amin Beheshti, Shahpar Yakhchi, Salman Mousaeirad, Seyed Mohssen Ghafari, Srinivasa Reddy Goluguri, and Mohammad Amin Edrisi. Towards cognitive recommender systems. *Algorithms*, 13(8):176, 2020b.

Elahi, M., Beheshti, A., & Goluguri, S. R. (2021). Recommender systems: Challenges and opportunities in the age of big data and artificial intelligence. *Data Science and Its Applications*, 15-39.

Blog, N. T. (2017, April 18). System architectures for personalization and recommendation. Medium.

<https://netflixtechblog.com/system-architectures-for-personalization-and-recommendation-e081aa94b5d8>