Sistema de recomendación con aplicación a recomendador de películas

Jhon Fredy Mercado
Dpto Ing. Telecomunicaciones
Universidad de Antioquia
jfredy.mercado@udea.edu.co

Keywords: Recommendation system, Deep learning, Collaborative Filtering

I. Contexto

El crecimiento explosivo en la cantidad de información digital disponible y el número de visitantes a Internet han creado un desafío potencial de sobrecarga de información que dificulta el acceso oportuno a elementos de interés en Internet. Los sistemas de recuperación de información, como Google, DevilFinder y Altavista, han resuelto parcialmente este problema Isinkaye, Folajimi, and Ojokoh (2015) [1], pero la priorización y la personalización (donde un sistema asigna el contenido disponible a los intereses y preferencias del usuario) de la información estaban ausentes. Esto ha aumentado la demanda de sistemas de recomendación más que nunca.

Los sistemas de recomendación son sistemas de filtrado de información que se ocupan del problema de la sobrecarga de información al filtrar fragmentos de información vital de una gran cantidad de información generada dinámicamente de acuerdo con las preferencias, el interés o el comportamiento observado del usuario.

El sistema de recomendación tiene la capacidad de predecir si un usuario en particular preferiría un artículo o no según el perfil del usuario. Los sistemas de recomendación son beneficiosos tanto para los proveedores de servicios como para los usuarios, ya que estos reducen los costos de encontrar y seleccionar artículos en sean de interés del usuario.

II. Objeto machine learning

Ya que para recomendar un contenido según las preferencias de un usuario se debe tener en cuenta tanto las características del usuario como las características del contenido que se tiene, para ello es necesario implementar tecnicas que modelen un filtrado colaborativo. Lo que se busca es implementar un modelo que seleccione el contenido que más se ajuste a las preferencias del usuario.

III. Dataset

Para el desarrollo de este proyecto se seleccionó el dataset The Movies Dataset proporcionado en kaggle, estos archivos contienen metadatos de 45.000 películas enumeradas en el conjunto de datos completo de MovieLens. Este conjunto de datos también tiene archivos que contienen 26 millones de calificaciones de 270.000 usuarios para las 45.000 películas. Las calificaciones están en una escala de 1 a 5 y se han obtenido del sitio web oficial de GroupLens.

El dataset tiene un total de 900 MB y contiene los siguientes archivos .csv

- **movies_metadata.csv**: Archivo principal de metadatos de películas, contiene información de 45000 películas incluyendo carteles, fondos, presupuesto, ingresos, fechas de lanzamiento, idiomas, países de producción y empresas.
- **keywords.csv**: Contiene las palabras clave de la trama de la película en forma de un objeto JSON en cadena.
- **credits.csv**: Contiene información sobre el reparto y equipo técnico de las películas en forma de objeto JSON en cadena.
- **links.csv**: Contiene los ID de TMDB e IMDB de las películas que aparecen en el conjunto de datos de MovieLens.
- **links_small.csv**: Contiene los ID de TMDB e IMDB de un subconjunto de 9.000 películas del conjunto de datos completo.
- **ratings_small.csv** El subconjunto de 100.000 calificaciones de 700 usuarios en 9.000 películas

IV. Métricas de desempeño

La calidad de un algoritmo de recomendación se puede evaluar utilizando diferentes tipos de medidas, que pueden ser de accuracy o de coverage. El accuracy es la fracción de recomendaciones correctas del total de recomendaciones posibles, mientras que la coverage mide la fracción de objetos en el espacio de búsqueda para los que el sistema puede proporcionar recomendaciones. Las métricas para medir el accuracy de los sistemas de filtrado de recomendaciones se dividen en métricas de precisión estadísticas y de soporte de decisiones.

Métricas de precisión estadística: Evaluar la precisión de una técnica de filtrado comparando las calificaciones pronosticadas directamente con la calificación real del usuario.

El error absoluto medio (MAE), el error cuadrático medio (RMSE) y la correlación se utilizan normalmente como métricas de precisión estadística.

MAE es el más popular y comúnmente utilizado, es una medida de la desviación de la recomendación del valor específico del usuario. Se calcula de la siguiente manera:

$$MAE = \frac{1}{N}\sum_{u,i}|p_{u,i} - r_{u,i}| \quad (1)$$

donde $p_{u,i}$ es la calificación prevista para el usuario u en el elemento i, $r_{u,i}$ es la calificación real y N es el número total de calificaciones en el conjunto de elementos. Cuanto menor sea el MAE, con mayor precisión el motor de recomendaciones predice las calificaciones de los usuarios. Además, el error cuadrático medio (RMSE) está dado por:

$$RMSE = \sqrt{\frac{1}{n}\sum_{u,i}(p_{u,i} - r_{u,i})^2} \quad (2)$$

El error cuadrático medio (RMSE) pone más énfasis en un error absoluto más grande y cuanto más bajo es el RMSE, mejor es la precisión de la recomendación.

Métricas de precisión de soporte de decisiones: Las características operativas del receptor (ROC) y la curva de recuperación de precisión (PRC), la precisión, la recuperación y la medida-F. Estas métricas ayudan a los usuarios a seleccionar elementos de muy alta calidad del conjunto de elementos disponibles. Las métricas ven el procedimiento de predicción como una operación binaria que distingue los elementos buenos de los que no lo son. Las curvas ROC son muy exitosas cuando se realizan evaluaciones integrales del rendimiento de algunos algoritmos específicos. La precisión es la fracción de elementos recomendados que son realmente relevantes para el usuario, mientras que el recuerdo puede definirse como la fracción de elementos relevantes que también forman parte del conjunto de elementos recomendados. Se calculan como:

$$Precision = \frac{ItemsCorrectamenteRecomendados}{TotalItemsRecomendados} \quad (3)$$

$$Recall = \frac{ItemsCorrectamenteRecomendados}{TotalItemsÚtilesRecomendados} \quad (4)$$

La medida-F definida a continuación ayuda a simplificar la precisión y la recuperación en una sola métrica. El valor resultante hace que la comparación entre algoritmos y entre conjuntos de datos sea muy simple y directa.

Otra métrica muy usada en los sistemas recomendadores de contenido es el Hit Ratio (HR) que es simplemente la fracción de usuarios para los que se incluye la respuesta correcta en la lista de recomendaciones de longitud L.

V. Referencias y resultados previos

Referente a los recomendadores de contenido se han tratado con diferentes enfoques y diferentes modelos, Isinkaye, Folajimi, and Ojokoh (2015) [1] resume las diferentes técnicas de sistemas recomendadores de contenido, como se ilustra en la Fig. 1, donde la selección de una o la combinación de varias

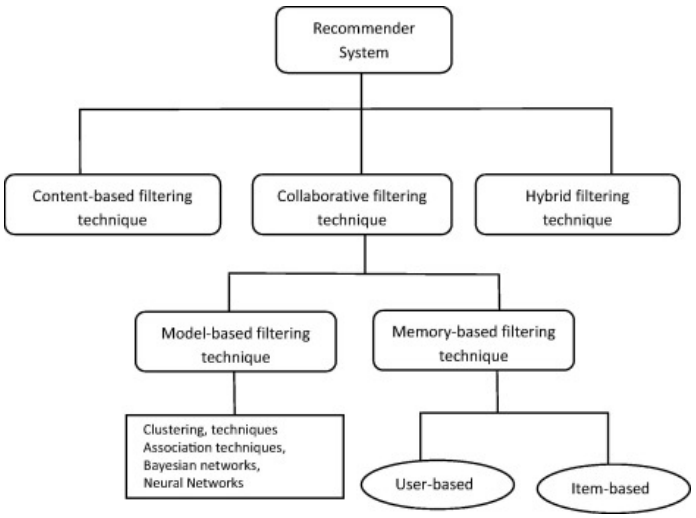

Fig. 1. Técnicas recomendador de contenidos.

de estas técnicas depende de la aplicación que se desea obtener y los datos disponibles.

Zamanzadeh Darban and Valipour (2022) [2] y Gan and Cui (2021) [3] son otros autores que proponen diferentes modelos más enfocados en técnicas de deep learning, proponiendo incluso un recomendador de contenido híbrido basado en grafos, comparándolo con diferentes modelos, teniendo un $RMSE$ de 0.833 una $presición$ de 0.792 y un $recall$ de 0.838 con un dataset de 1M de calificaciones de películas

References

[1] F. Isinkaye, Y. Folajimi, and B. Ojokoh, "Recommendation systems: Principles, methods and evaluation," *Egyptian Informatics Journal*, vol. 16, no. 3, pp. 261–273, 2015, ISSN: 1110-8665. DOI: https://doi.org/10.1016/j.eij.2015.06.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110866515000341.

[2] Z. Zamanzadeh Darban and M. H. Valipour, "Ghrs: Graph-based hybrid recommendation system with application to movie recommendation," *Expert Systems with Applications*, vol. 200, p. 116 850, 2022, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.116850. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422003025.

[3] M. Gan and H. Cui, "Exploring user movie interest space: A deep learning based dynamic recommendation model," *Expert Systems with Applications*, vol. 173, p. 114 695, 2021, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.114695. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421001366.