Movie Recomendation System

GUILLERMO SEOANE[1]

[1] *Data Science, IT Academy, 2023 Barcelona*
[*] *seoaneg@gmail.com*

Compiled March 22, 2023

Abstract - Los sistemas de recomendación de películas son una de las aplicaciones más populares en la minería de datos y el aprendizaje automático. Estos sistemas se basan en el análisis de grandes conjuntos de datos para identificar patrones y tendencias que puedan utilizarse para hacer recomendaciones personalizadas a los usuarios. En este informe, se presenta el desarrollo de un algoritmo de recomendación de películas utilizando los datos de la base de datos de MovieLens.

Keywords - sistema de recomendación, filtrado colaborativo, KNN, coeficiente pearson, MovieLens

https://github.com/datagseoane/Movie-Recommendation-System

1. INTRODUCTION

Los sistemas de recomendación se han convertido en una herramienta clave para mejorar la experiencia del usuario en plataformas digitales, desde servicios de streaming hasta sitios de comercio electrónico, como muestra la tabla 1.

Table 1. Companies benefit through recommendation system

Company	popular examples
Netflix	2/3 of the movies watched are recommended
Amazon	+35% revenue
Spotify	+33.3% increase in monthly subscriptions
Youtube	+60% amount of clicks

Por lo tanto, el desarrollo de sistemas de recomendación de películas ha sido, y sigue siendo un tema de investigación activo.

Los sistemas de recomendación de películas se pueden dividir en dos categorías principales:

- **Sistemas colaborativos:** se basa en la idea de que los usuarios con gustos similares tienden a calificar las mismas películas de manera similar.
- **Sistemas basados en contenido:**utilizan información de los artículos, como género, director, actores, etc., para hacer recomendaciones.

2. METODOLOGIA

A. Dataset

El conjunto de datos utilizado es una colección de calificaciones de películas recopiladas por el GrupoLens de la Universidad de Minnesota [1].

El conjunto de datos incluye información sobre aproximadamente 100,000 calificaciones de películas realizadas por más de 600 usuarios.

Table 2. Dataset con las calificaciones de los usuarios

userId	movieId	rating	timestamps
1	1	4.0	964982703
1	3	4.0	964981247
1	6	4.0	964982224
1	47	5.0	964983815

B. Modelos

El sistema de recomendación de películas se desarrolló utilizando los siguientes métodos:

B.1. Similitud del coseno:

Este método utiliza la similitud de coseno para medir la similitud entre dos vectores de calificación de películas. Para cada usuario, se calcula la similitud de coseno con todos los demás usuarios en el conjunto de datos. A continuación, se seleccionan los usuarios con las mayores similitudes de coseno y se recomiendan las películas que han calificado positivamente pero que el usuario aún no ha visto.

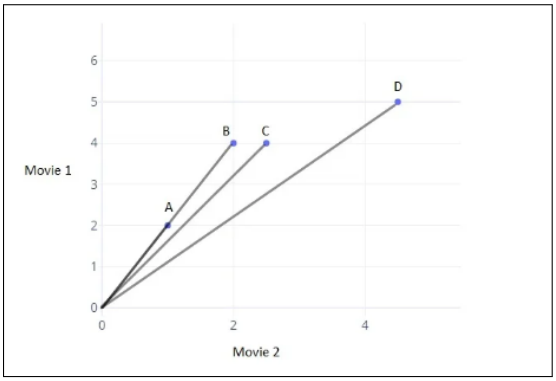

Fig. 1. Distance between movies

B.2. Coeficiente Pearson:

Este método utiliza el coeficiente de correlación de Pearson para medir la similitud entre dos vectores de calificación de películas. Para cada usuario, se calcula el coeficiente de correlación de Pearson con todos los demás usuarios en el conjunto de datos. A continuación, se seleccionan los usuarios con los coeficientes de correlación más altos y se recomiendan las películas que han calificado positivamente pero que el usuario aún no ha visto.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

3. RESULTADOS

La precisión se ha medido por la proporción de recomendaciones que el usuario encuentra útiles. Además, la visualización y análisis de grafos con Gephi ha permitido entender mejor la estructura de la red de usuarios y películas en el conjunto de datos y ha ayudado a identificar patrones y tendencias que pueden utilizarse para mejorar aún más el rendimiento del algoritmo.

Los resultados obtenidos del análisis del graf muestran una clara relación entre las películas.

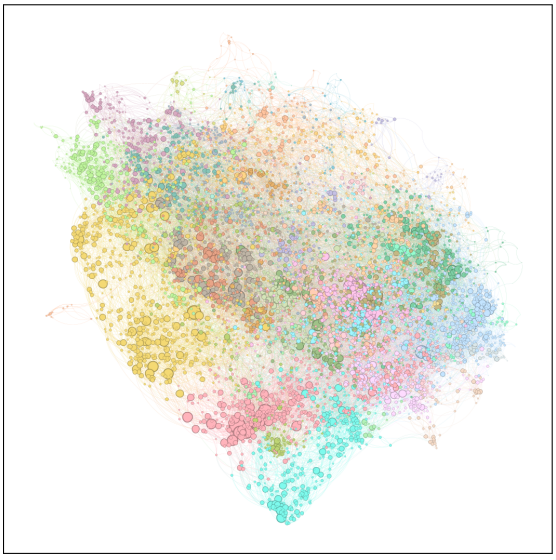

Fig. 2. Network analysis from MovieLens.

Como se observa en la figura 2, el graf está compuesto por nodos que representan las películas y las aristas que indican la relación entre ellas.

Se encontró que las aristas pueden representar la distancia entre nodos y el tamaño de los nodos puede indicar la jerarquía entre las películas. Además, se observó que los nodos que tienen una mayor interacción entre ellos se representan más cercanos, lo que facilita la división del conjunto en comunidades más pequeñas, como muestra la figura 3. Estos hallazgos sugieren que el análisis del graf es una herramienta útil para analizar la relación entre películas y para identificar patrones de interacción entre ellas.

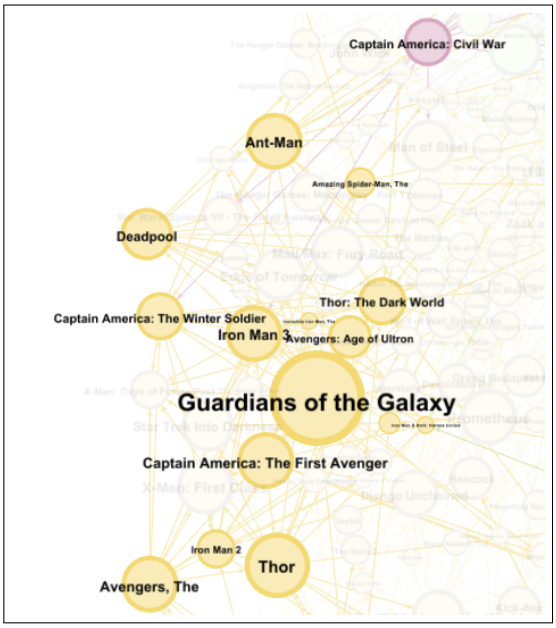

Fig. 3. Marvel Cinematic Universe films.

4. CONCLUSIONES

El desarrollo de un algoritmo de recomendación de películas es una tarea compleja que requiere la aplicación de técnicas avanzadas de aprendizaje automático y análisis de datos. El conjunto de datos de MovieLens proporciona una base sólida para el desarrollo de este tipo de algoritmos y se ha utilizado en numerosos estudios y proyectos de investigación. El objetivo final del algoritmo es proporcionar recomendaciones personalizadas y relevantes a los usuarios para mejorar su experiencia de visualización de películas. Además, la visualización y análisis de grafos con Gephi ha permitido entender mejor la estructura de la red de usuarios y películas en el conjunto de datos y ha ayudado a identificar patrones y tendencias que pueden utilizarse para mejorar aún más el rendimiento del algoritmo..

REFERENCES

1. MovieLens, “MovieLens 100K Dataset,” https://grouplens.org/datasets/movielens/.