Tecnológico de Monterrey

Análisis de grandes volúmenes de datos (Gpo 10)

Equipo 24

Sistemas de Recomendación - Comparación y limitaciones de algoritmos de procesamiento de *Big Data*

Luis Salomón Flores Ugalde - A00817435
Oscar Israel Lerma Franco - A01380817
Alejandro Guzmán Chávez - A01795398

09/06/2024

En esta entrega es necesario realizar un reporte donde se enlisten los siguientes aspectos:

- **Revisión** del alcance y objetivos del proyecto. Discute y analiza con tus compañeros de equipo cualquier modificación requerida para ajustes al alcance u objetivos del proyecto. Documenta los cambios si los hay, en caso contrario, escribe sin cambios en este rubro.
- Realiza la **comparación** de diferentes algoritmos de recomendación de las actividades 4.2 y 6.2, en términos de rendimiento y escalabilidad. Crea una tabla que demuestre la evidencia de la comparación.

Criterio	Valor
Portada con datos completos de los integrantes	5
Revisión del alcance y objetivos del proyecto.	30
Realiza la **comparación** de diferentes algoritmos de recomendación de las actividades 4.2 y 6.2, en términos de rendimiento y escalabilidad. Crea una tabla que demuestre la evidencia de la comparación.	65
Total	100

Alcance y objetivos de proyecto

Alcance de proyecto

Desde el inicio del proyecto, se expresó el interés de proporcionar a los usuarios recomendaciones *relevantes* de videojuegos, en base de actividad (o en el caso actual, reseñas). La necesidad de conectar al usuario con contenido relevante es cada vez más requerido por desarrolladores, y distribuidores. (véase "indipocalipse[1]")

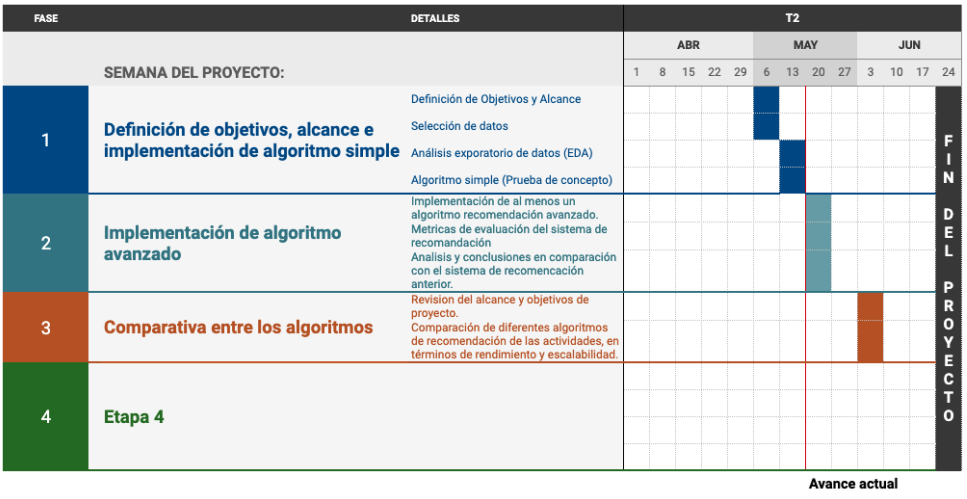
- *No se han realizado cambios de los objetivos generales del proyecto:*

Objetivos

- Obtener un sistema de recomendación efectivo para la recomendación y exploración de artículos para jugadores.
- El enfoque del sistema en los avances debería mejorar los resultados actuales en cuanto a la Precisión@K y MAP@K

Cronograma del Avance 3

TÍTULO DEL PROYECTO	Implementación de modelo de recomendación	**EMPRESA**	INSTITUTO TECNOLÓGICO DE MONTERREY
RESPONSABLE DEL PROYECTO	Equipo 24	**FECHA**	09/06/24



[1] "Indiepocalipse": Nombre popular dado a eras de pocas ventas de videojuegos independientes.

Comparación de algoritmos:

	Metric	Cosine Method	SVD Method
0	Precision@K	0.176965	0.000000
1	Recall@K	0.647169	0.000000
2	MAP@K	0.307603	0.000000
3	NDCG@K	0.411096	0.000000
4	MRR@K	0.388815	0.000000
5	Tiempo Entrenamiento	2.993277	8.410636
6	Tiempo Recomendación	79.673120	47.635721

Entre las métricas de precisión utilizadas:

- **Precisión@K**: Mide el porcentaje de ítems recomendados entre los primeros K elementos relevantes. Cuando se prefiere calidad de recomendaciones a su cantidad.
- **Recall@K**: Mide el porcentaje de elementos relevantes encontrados entre los primeros K recomendados. Evalúa cuántos elementos relevantes son descubiertos por el sistema. Se calcula como el número de ítems relevantes en las primeras K posiciones dividido por el número total de ítems relevantes.
- **MAP@K** (Mean Average Precision at K): Promedio de las precisiones calculadas en cada posición de un ítem relevante hasta K, promediado sobre todos los usuarios o consultas. Considera la posición de los ítems relevantes y su cantidad, dando una medida del rendimiento de la clasificación.
- **NDCG@K** (Normalized Discounted Cumulative Gain at K): Evalúa la clasificación de los ítems recomendados dando más importancia a los aparecen en las posiciones superiores de la lista de recomendaciones.
- **MRR@K** (Mean Reciprocal Rank at K): Es el promedio del recíproco de la posición del primer ítem relevante en la lista de recomendaciones, calculado sobre todas las consultas. Cuando sólo importa la posición del primer ítem relevante.

Resultados y análisis

Con los valores obtenidos anteriormente, observamos:

SVD:

- Un claro problema con el método SVD que estamos usando. Cierto que también vimos el RMSE de ambos métodos y el SVD daba un resultado de entre 0.12 y 0.35 dependiendo de un filtrado en las recomendaciones. Pero hablando exclusivamente de nuestras métricas. Creemos que en realidad tiene sentido que tengan 0. Esto, por el *sparsity* de los datos.
- Los datos usados son muy dispersos cuando solo tomamos en cuenta la columna de “recommend” qué es lo que justo ahora toma en cuenta el método SVD. Un usuario puede tener 20+ juegos en su librería pero había ‘recomendado’ 1/20 y es todo lo que se toma en cuenta al momento. Pero esto trae consigo un problema que podemos ver rápidamente, realmente el único dato que podríamos extraer para decir nosotros que a un usuario “le gusto” algo es el tiempo. Y de esto podríamos arbitrariamente decidir que si un usuario jugó más de X horas de juego pues lo consideramos como positivo. Pero esto requiere un análisis diferente y nuestra fuente de datos al final no cuenta con estos datos de librerías de forma completa, solo tenemos acceso a los datos Australianos.
- Expandir el algoritmo a que tome la información de las librerías sería el siguiente paso lógico para mejorar el abismal resultado del algoritmo pero esto implicaría trabajar más sobre el modelo SVD. Pero no podríamos experimentar con información fuera de Australia de esta forma.
- Aunque tenga espacio para escalar, no creemos que sea suficiente para justificarlo. En especial cuando pensamos en la evidencia presente de *sparsity*.

Similaridad del Coseno:

- Este algoritmo mide la similitud entre dos vectores, Es comúnmente utilizado en filtros colaborativos basados en memoria.
- Toma como base los géneros y los tags de los juegos. Esto permite comparar los vectores de dos juegos para encontrar los más similares a cualquier otro de ellos.
- Si bien las métricas no son tan buenas, se puede trabajar en mejorarlas para incrementar la precisión de los artículos recomendados relevantes. Al momento la precisión es de 17%, quiere

decir que el 17% de recomendaciones fueron relevantes para el usuario.

- A pesar de que el algoritmo no interpreta las relaciones complejas entre los registros de los usuarios, se puede implementar el algoritmo utilizando más datos del comportamiento del usuario que puedan describir su comportamiento. Por ejemplo, recomendaciones de tipo: “porque jugaste X te recomendamos Y” basándonos en el tiempo de juego del usuario en el juego X se puede hacer uso del algoritmo de coseno para buscar juegos similares a ese. Otras relaciones que se pueden utilizar son: “porque compraste X te recomendamos Y” o “porque valoraste X con un buen rating, te recomendamos Y”. Esto dependería del contexto en el que se plantea implementar el algoritmo para enviar las recomendaciones al usuario.
- Otra métrica relevante es la Mean Reciprocal Rank, que indica que la primera recomendación relevante está en la posición 1/0.38 = 2.63. En la posición 2.6 se encuentra el primer item relevante para el usuario. Hablando de una plataforma de videojuegos, el primer item relevante está muy cerca de la vista del usuario. También se puede mejorar pero no es una métrica mala de momento.

Tabla comparativa

Aspecto	**SVD**	**Coseno**
Principio	Descomposición matricial de usuarios-item	Similitud del coseno entre vectores
Ventajas	Maneja datos escasos, reducción dimensional	Simplicidad, escalabilidad
Desventajas	Costoso computacionalmente, re-calculación	No captura relaciones complejas
Casos de uso	Plataformas de streaming, e-commerce	Recomendaciones de artículos, filtrado colaborativo simple

Conclusión

En esta etapa del proyecto hemos comparado los dos algoritmos de recomendación: SVD (Descomposición de Valores Singulares) y Similaridad del Coseno. Ambos métodos fueron evaluados utilizando métricas de precisión, recall, MAP, NDCG y MRR para determinar su eficacia de recomendación de videojuegos basados en reseñas y actividad de los usuarios.

El sistema de recomendación *item-item* calculado con similitud de coseno, mostró indicios de progreso relevante en comparación con el sistema de recomendación anterior (SVD simple), mostrándose como la mejor opción para nuestras necesidades actuales. mientras que SVD podría ser explorado más a fondo con un conjunto de datos más completo.

La implementación de mejoras y la incursión de más datos podrían incrementar la eficacia de ambos métodos a futuro. En la siguiente fase del proyecto se espera trabajar en la mejora de las métricas precisión y MAP para el algoritmo de similitud de coseno.

Github

Notebook directo:
https://github.com/VF1Gimure/MNA_GVD24/blob/main/Proyecto_Avance_3_24.ipynb

Carpeta del proyecto: https://github.com/VF1Gimure/MNA_GVD24/tree/main

Bibliografía

1. Sun, J., Gan, W., Chen, Z., Li, J., & Yu, P. S. (2022). Big data meets metaverse: A survey. arXiv preprint arXiv:2210.16282. https://arxiv.org/pdf/2210.16282Links to an external site.
2. Kang W., McAuley J. (*ICDM*, 2018) **Self-attentive sequential recommendation.** UC San Diego pdf