Minería de datos para grandes volúmenes de información

Proyecto: Sistema de recomendación de películas

Giovanny Gómez Convers
Sun Yin Huang Huang

2023-2

---

1. Pregunta de investigación y objetivos

En la actualidad los sistemas de recomendación son una herramienta indispensable para ayudar a los usuarios a descubrir elementos de interés que pueden ser imperceptibles o inapreciables a primera vista. Haciendo uso de diferentes técnicas tales como el filtro colaborativo, que tiene relación con información dada por otros usuarios, o, el filtro de contenido, que pondera los atributos del producto o servicio de interés, encontramos que es importante determinar el impacto que tienen en primer lugar, sobre la toma de decisiones de compra, en segundo lugar, sobre la mejora en la experiencia de usuario y finalmente sobre la fidelización de clientes que se espera regresen y sigan comprando. En particular, abordaremos la influencia y precisión que tienen estos sistemas de recomendación en la elección de preferencias cinematográficas en plataformas de streaming de películas.

Objetivo principal

Desarrollar un sistema de recomendación basados en filtros colaborativo y de contenido y evaluar el impacto sobre la selección de películas en plataformas de streaming que tienen los sistemas de recomendación y cómo influye en la exploración de contenido por parte de los usuarios.

Objetivos específicos

- Evaluar precisión de los principales algoritmos de recomendación en la selección de películas en plataformas de streaming.
- Examinar que características o hiperparámetros tienen una mayor influencia sobre los algoritmos de recomendación.
- Identificar posibles sesgos en los algoritmos de recomendación y evaluar su impacto en la equidad y toma de decisiones que realizan los usuarios.

2. Revisión de literatura

Un sistema de recomendación tiene como objetivo anticipar y comprender los intereses de los usuarios, basándose en el análisis de su historial de comportamiento. Estos datos históricos representan un componente fundamental en el proceso de personalización, ya que proporcionan información valiosa sobre las preferencias y elecciones previas del usuario. La recopilación y el análisis de esta información permiten al sistema identificar patrones, tendencias y afinidades que son esenciales para ofrecer recomendaciones precisas y relevantes (Melville y Sindhwani, 2010).

Además, es importante destacar que los datos pasados no solo reflejan las acciones pasadas del usuario, sino que también son el producto de las decisiones tomadas por él mismo. En otras palabras, cada vez que un usuario interactúa con el sistema de recomendación, está tomando decisiones que afectan su experiencia futura. Estas decisiones pueden incluir la selección de productos o servicios de una lista de recomendaciones, la especificación de preferencias en términos de características o la realización de consultas de búsqueda específicas. Como consecuencia de estas interacciones, los usuarios experimentan una personalización significativa en su experiencia. Esto se traduce en el hecho de que, en lugar de ver recomendaciones genéricas, cada usuario recibe listas de recomendaciones adaptadas a sus gustos, preferencias y necesidades específicas (Jannach et al., 2010).

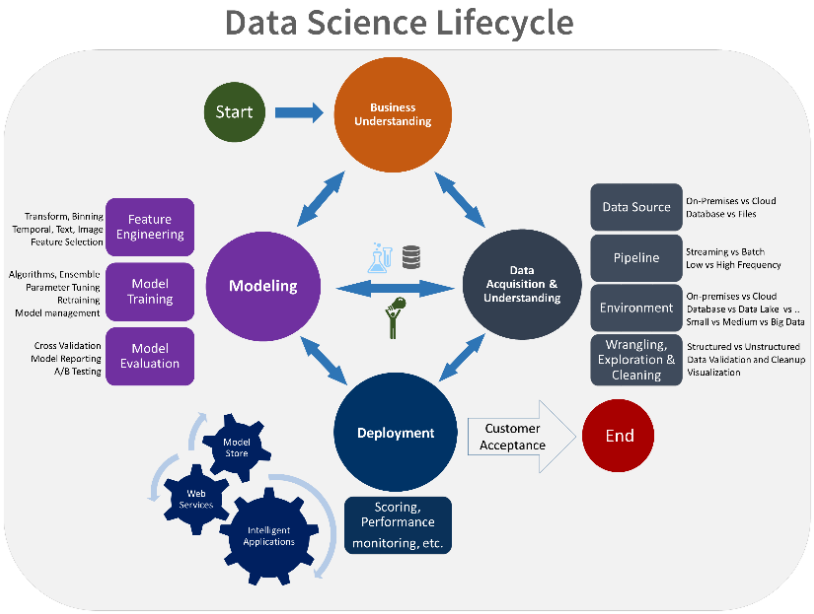
Los sistemas de recomendación se han convertido en un área de investigación importante desde la aparición del primer artículo de investigación sobre filtrado colaborativo a mediados de la década de 1990 (Resnick et al., 1994, Shardanand y Maes, 1995) y actualmente se está convirtiendo en uno de los métodos más importantes para proporcionar documentos, mercancías, y cooperadores para responder a las necesidades de los usuarios en la provisión de información, comercio y servicios que sean para la sociedad (Melville y Sindhwani, 2010).

La clave para una recomendación exitosa es tener un mecanismo de recomendación poderoso que pueda vincular los artículos del producto con las preferencias del usuario. Estos mecanismos suelen caer en las áreas de recuperación de información o minería de datos (Baeza-Yates y Ribeiro-Neto, 1999). Teniendo en cuenta el tipo de datos de entrada y los métodos utilizados, en la literatura podemos encontrar tres enfoques: filtrado basado en demografía, filtrado colaborativo y filtrado basado en contenido (Pazzani, 1999).

El filtrado basado en datos demográficos analiza las preferencias del usuario y hace recomendaciones basadas en la información demográfica del usuario como, por ejemplo, la edad, el sexo y los niveles de ingresos. El filtrado colaborativo determina las preferencias del usuario y recomienda en función de los estereotipos encontrados del usuario. Es decir, el comportamiento de un usuario se predice basándose en evidencia que ha mostrado similitud con otros usuarios. Y el filtrado basado en contenido hace recomendaciones basadas en ciertos atributos del producto como, por ejemplo, el atributo de una película puede ser el director, los actores y el productor. No obstante, también podemos encontrar métodos híbridos que son una combinación de recomendaciones basadas en contenido y colaboración (Melville y Sindhwani, 2010).

3. Metodología de investigación

Mas información en: ¿Qué es el Proceso de ciencia de datos en equipo (TDSP)? - Azure Architecture Center | Microsoft Learn



4. Análisis de los datos

Descripción de los datos

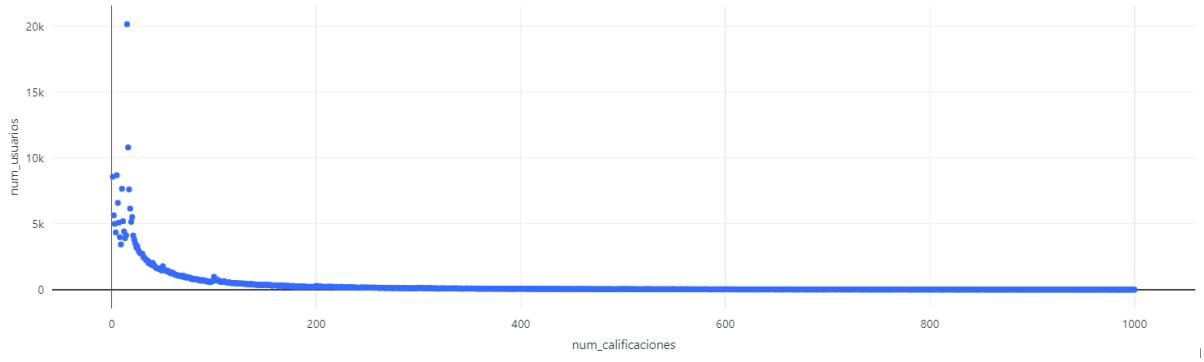
El dataset contiene información de las calificaciones de estrellas (0-5) y etiquetado de texto libre de MovieLens. Abarca una amplia gama de géneros cinematográficos y épocas, lo que lo convierte en una herramienta valiosa para la investigación y el desarrollo de sistemas de recomendación.

MovieLens es una plataforma que ofrece un servicio de recomendación de películas dirigido por el GroupLens, un laboratorio de investigación de la Universidad de Minnesota que desarrolla nuevas herramientas e interfaces experimentales para la exploración y recomendación de datos. MovieLens no es comercial, no contiene publicidad y sus datos son de acceso libre.

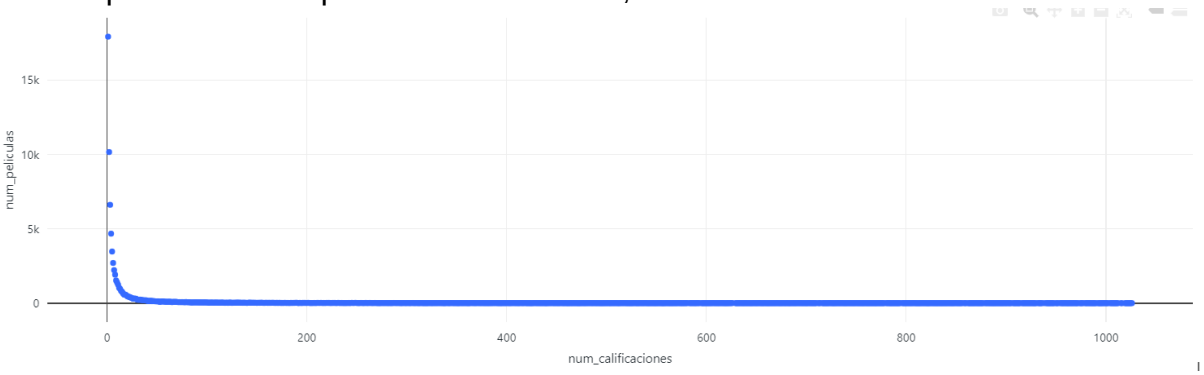
El dataset tiene 33.832.162 millones de calificaciones provenientes de 330.975 usuarios que participaron entre el 9 de enero de 1995 y el 20 de julio de 2023 para un total de 83.239 películas evaluadas.

Análisis exploratorio

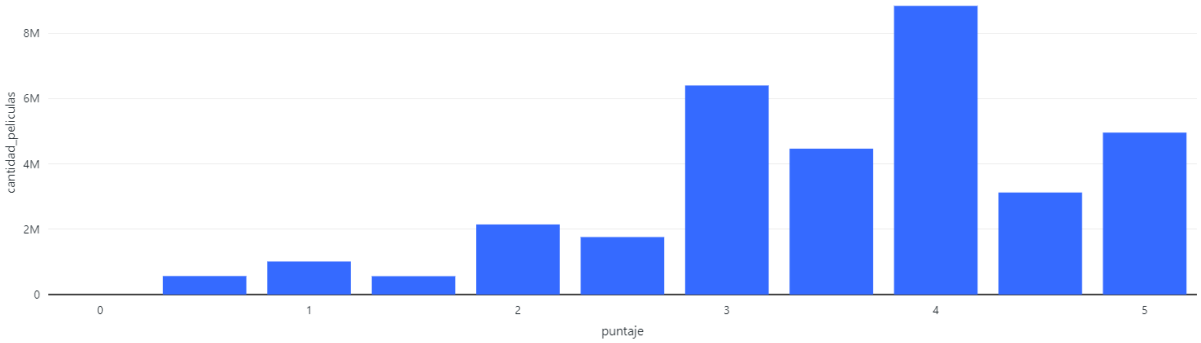
En promedio, cada usuario ha calificado 102,2 películas.



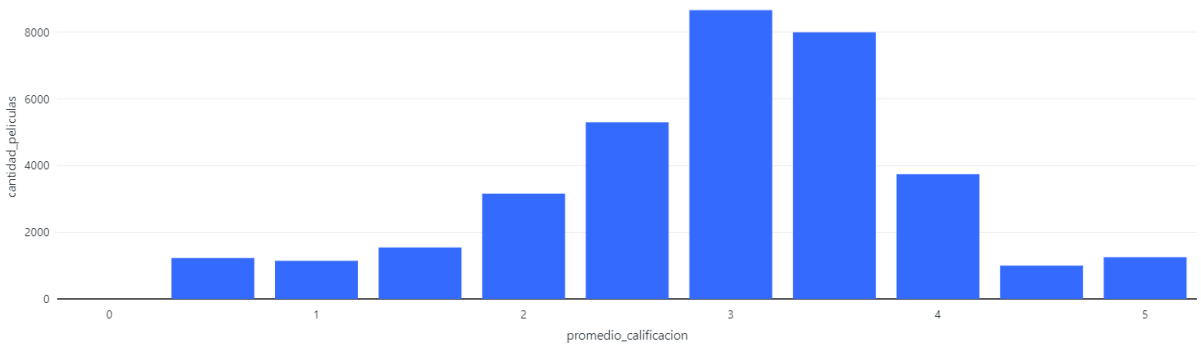
Y cada película tiene aproximadamente 406,5 calificaciones.



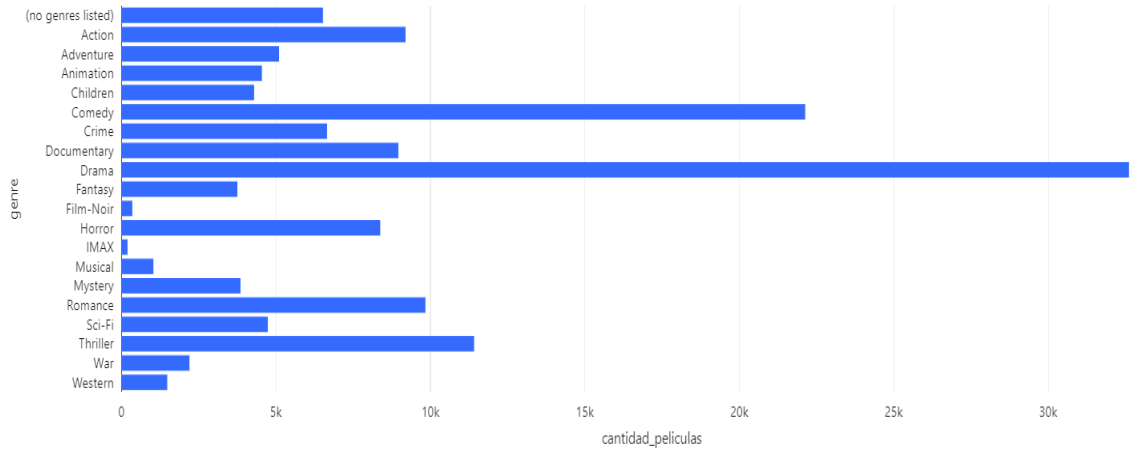
Encontramos que la mayoría de las calificaciones se encuentran por encima de las 3 estrellas.



Y que gran parte de las películas se encuentran en el rango medio.



Finalmente, las películas son etiquetadas principalmente en los géneros de drama, comedia y thriller.



Los datos fueron cargados al metastore hive de Databricks y desde allí son consultados por los modelos y análisis de datos:

Ratings

default.ratings

Propietario: No se ha establecido Tamaño: 213.9MiB, 1 archivo Última actualización: hace 2 meses

Created by the file upload UI

Columnas | Datos de muestra | Detalles | Permisos | Historial

Filtrar columnas...

Columna	Tipo	Com
userId	bigint	
movieId	bigint	
rating	double	
timestamp	bigint	

Movies

default ›

default.movies

Propietario: No se ha establecido Tamaño: 2.1MiB, 1 archivo Última actualización: hace 2 meses

Created by the file upload UI

Columnas | Datos de muestra | Detalles | Permisos | Historial

Filtrar columnas...

Columna	Tipo	C
movieId	bigint	
title	string	
genres	string	

5. Uso de herramientas de Big Data

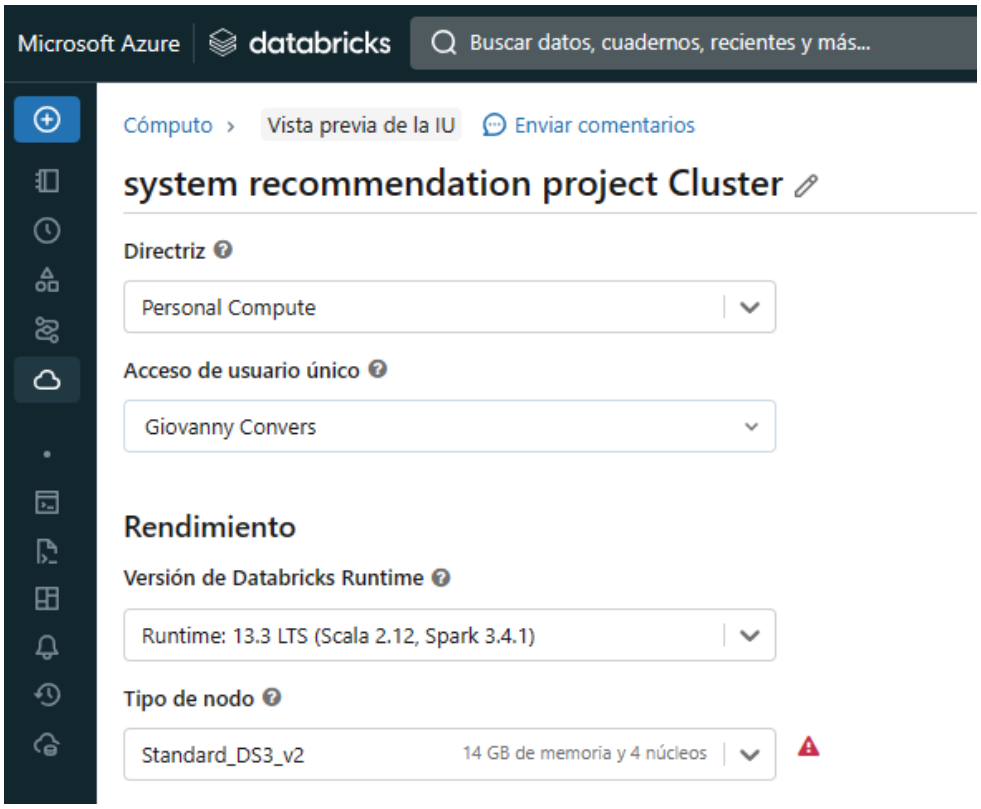
Es fundamental tener en cuenta que, en el contexto del desarrollo de este trabajo, el manejo de una gran cantidad de información fue un aspecto crítico para el éxito del sistema de recomendación. Para abordar este desafío, se recurrió al uso de herramientas claves como Databricks y Apache Spark.

Databricks es una plataforma de análisis de datos en la nube que se basa en el popular motor de procesamiento de datos Apache Spark. Databricks proporciona un entorno unificado que permite a los equipos de datos colaborar en la ingesta, procesamiento, análisis y visualización de datos a gran escala. Esta plataforma simplifica la gestión de flujos de trabajo de datos, lo que la hace especialmente valiosa para proyectos que involucran grandes volúmenes de datos, como el desarrollo de sistemas de recomendación.

Por otro lado, Apache Spark es un potente marco de procesamiento de datos distribuido que se utiliza para realizar operaciones de procesamiento en paralelo en conjuntos de datos masivos. Spark permite el procesamiento de datos en tiempo real y por lotes, y es altamente escalable. Su capacidad para realizar análisis y transformaciones de datos a gran velocidad lo convierte en una elección ideal para proyectos que implican la manipulación y análisis de datos a gran escala.

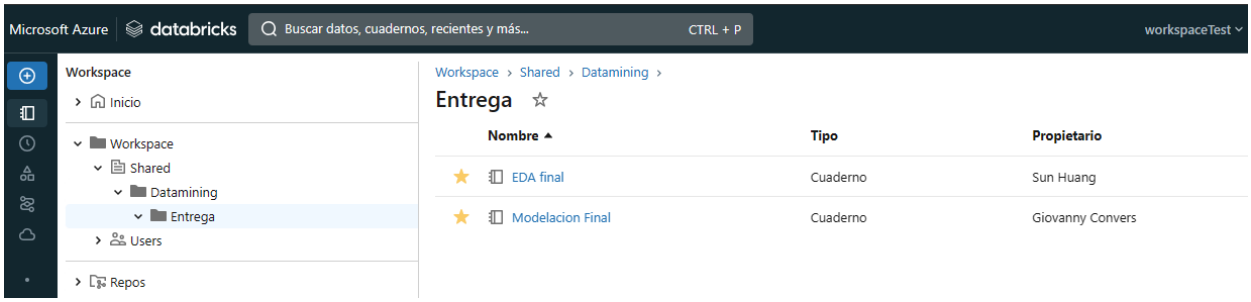
En el notebook se puede encontrar más detalle sobre el uso de estas herramientas.

Se aprovisionó un cluster con la siguiente configuración:



Url: https://adb-1526615590818240.0.azuredatabricks.net/

Ubicación de notebooks



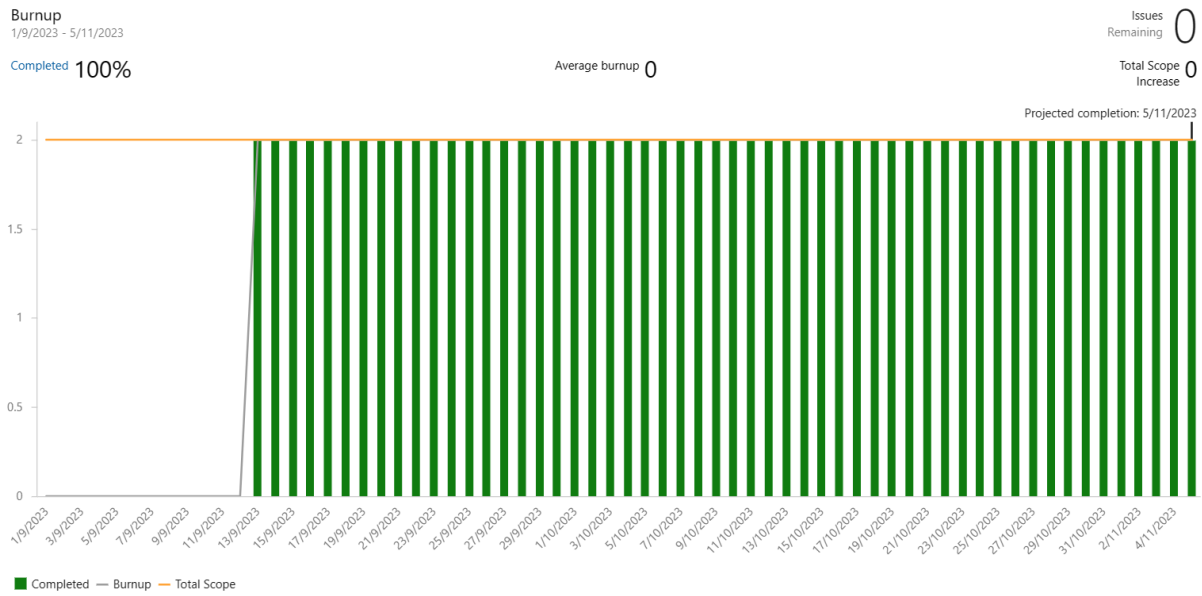
6. Conclusiones y trabajo futuro

https://github.com/giovy79/BigDataMining

Algoritmos	Resultados
Filtro colaborativo basado en memoria (vecinos cercanos)	Ejecución después de 58,60 minutos sin respuesta.
Filtro colaborativo basado en modelos	RMSE en datos de testeo 0.8219
Filtro contenido	RMSE en datos de testeo 0.8072

7. Ejecución del plan

https://dev.azure.com/EafitMCDA2023/BigDataMining



En la planeación realizada se definieron sprints de dos semanas en los cuales los correspondientes al mes de septiembre no se ejecutaron las actividades al principio, lo cual nos generó mover las actividades durante los siguientes sprints y no contar con holgura. En ese sentido, observamos que las actividades se cumplieron, pero nos generó un desfase al final.

8. Implicaciones éticas

Las principales implicaciones tienen que ver con la manipulación de las decisiones de los usuarios al priorizar contenido polarizado, en el cual, los usuarios eviten una mirada crítica o cuestionen los resultados ofrecidos, tomando así una posición consumista e irreflexiva frente a las recomendaciones sugeridas. En este escenario, se puede presentar un gran riesgo de manipulación comportamental e inducir a selecciones sectarias que si bien priorizan las preferencias pueden fomentar la discriminación.

9. Aspectos legales y comerciales

No se consideran implicaciones o consecuencias legales dado que el conjunto de datos es de libre acceso. Con respecto a los resultados obtenidos, se observa viabilidad para realizar un uso comercial de los modelos seleccionados y de los indicadores o selección de características principales que mayor incidencia tienen en las recomendaciones realizadas.

Referencias bibliográficas

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). Recommender systems: an introduction, Cambridge University Press

Melville, P., & Sindhwani, V. (2010). Recommender systems. Encyclopedia of machine learning, 1, 829-838.

Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. Artificial intelligence review, 13, 393-408.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, NC, pp. 175-186.

Riedl, J., Jameson, A., & Konstan, J. (2004). AI Techniques for Personalized Recommendation. Proposal for a full-day AAAI tutorial.

Shardanand, U., & Maes, P. (1995, May). Social information filtering: Algorithms for automating “word of mouth”. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 210-217).