

Pontificia Universidad Javeriana Cali
Facultad de Ingeniería y Ciencias
Ingeniería de Sistemas y Computación
Anteproyecto de Grado

Sistemas de Recomendación Basados en la Inferencia de Redes Sociales

Santiago Uribe Pastás

Director: Dr. Jorge Finke Ortiz
Co-director: Dr. Carlos Ernesto Ramírez Ovalle

Octubre 9 de 2021



Santiago de Cali, Octubre 9 de 2021.

Señores

Pontificia Universidad Javeriana Cali.

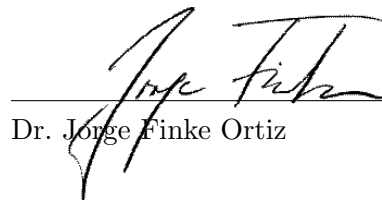
Dr. Gerardo Mauricio Sarria Montemiranda

Director Carrera de Ingeniería de Sistemas y Computación
Cali.

Cordial Saludo.

Por medio de la presente me permito informarle que el estudiante de Matemáticas Aplicadas Santiago Uribe Pastás (cod: 8925546) trabaja, bajo nuestra dirección en el proyecto de grado titulado “Sistemas de Recomendación Basados en la Inferencia de Redes Sociales”.

Atentamente,



Dr. Jorge Finke Ortiz



Dr. Carlos Ernesto Ramírez Ovalle

Santiago de Cali, Octubre 9 de 2021.

Señores

Pontificia Universidad Javeriana Cali.

Dr. Gerardo Mauricio Sarria Montemiranda

Director Carrera de Ingeniería de Sistemas y Computación
Cali.

Cordial Saludo.

Me permito presentar a su consideración el anteproyecto de grado titulado “Sistemas de Recomendación Basados en la Inferencia de Redes Sociales” con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el proyecto de grado y posteriormente optar al título de Matemático Aplicado

Al firmar aquí, doy fe que entiendo y conozco las directrices para la presentación de trabajos de grado de la Facultad de Ingeniería aprobadas el 26 de Noviembre de 2009, donde se establecen los plazos y normas para el desarrollo del anteproyecto y del trabajo de grado.

Atentamente,

Santiago Uribe P.
Santiago Uribe Pastás
Código: 8925546

Resumen

Los sistemas de recomendación son una parte vital para empresas con una activa participación en la web. Dichas empresas requieren de estrategias que les permitan apalancarse en las calificaciones de productos por parte de usuarios, para poder brindar futuras recomendaciones a otros usuarios. En la última década se han desarrollado diversos algoritmos para la recomendación de películas, siendo el algoritmo llamado *Matrix Factorization* uno de los más populares. El enfoque de esta tesis es evaluar el desempeño de este algoritmo de recomendación en escenarios en los que se pueden inferir redes sociales subyacentes, que caracterizan cierto tipo de interacción entre usuarios. En particular se utiliza el conjunto de datos *MovieLens*, el cual consta de aproximadamente cien mil valoraciones por parte de 671 usuarios sobre 9066 películas, durante el periodo del 29 de marzo de 1996 al 24 de septiembre de 2018. En esta tesis se investigará diferentes métodos para la inferencia de redes sociales, así como diferentes medidas topológicas sobre estas redes inferidas. Y con base en esta investigación, escoger y analizar un algoritmo que permita entender cómo el uso de la información de las redes sociales afecta el desempeño del algoritmo *Matrix Factorization*.

Palabras Clave: Sistemas de Recomendación, Redes Sociales Inferidas, *Matrix Factorization*, Grafos.

Índice general

1. Descripción del Problema	11
1.1. Planteamiento del Problema	11
1.1.1. Formulación	12
1.1.2. Sistematización	12
1.2. Objetivos	12
1.2.1. Objetivo General	12
1.2.2. Objetivos Específicos	12
1.3. Justificación	12
1.4. Delimitaciones y Alcances	13
1.4.1. Entregables	13
2. Desarrollo del Proyecto	15
2.1. Marco de Referencia	15
2.1.1. Áreas Temáticas	15
2.1.2. Marco Teórico	15
2.2. Medidas Topológicas de Redes Sociales	18
2.3. Trabajos Relacionados	19
2.4. Métodos para Inferir Redes Sociales	21
2.5. Metodología	22
2.5.1. Tipo de Estudio	22
2.5.2. Actividades	22
2.6. Resultados Esperados	23
2.7. Cronograma	23
2.8. Recursos	23
2.8.1. Recursos Humanos	23
2.8.2. Recursos Técnicos	24
2.8.3. Presupuesto	24
Bibliografía	25

Introducción

Con el constante aumento del volumen de la información en línea, los sistemas de recomendación (SR) han sido una herramienta utilizada por un amplio conjunto de empresas, para facilitar el encontrar información relevante que satisfaga las necesidades e intereses de usuarios, y así incrementar sus niveles de satisfacción en la interacción con diferentes marcas y servicios.

Diferentes implementaciones de SR intentan sugerir todo tipo de productos para los usuarios. Hoy en día se encuentran en la mayoría de las plataformas de empresas con alta participación en la web. Se utilizan para recomendar música en Spotify, productos varios por Amazon, vídeos en YouTube, y hasta artículos de investigación a la comunidad científica [1]. En el caso del material audiovisual, plataformas de *streaming* (como Netflix, Amazon Prime Video y HBO) aprovechan una variedad de conjuntos de datos y técnicas de *Machine Learning* para generar las recomendaciones [2].

En los últimos años, se han generado diferentes algoritmos para la recomendación de películas, los cuales han demostrado una alta efectividad [3]. El filtrado colaborativo es uno de los enfoques más usados en SR. Este tipo de algoritmo parte de la base que si los usuarios han valorado artículos de forma similar en el pasado, es probable que valoren artículos semejantes de forma similar en el futuro. Aunque los modelos de filtrado colaborativo generan resultados prometedores en diferentes escenarios, no explotan de forma explícita patrones ocultos en los datos de redes sociales que podrían agregar valor a la predicción de las valoraciones de los usuarios.

Inferir las redes sociales subyacentes a un conjunto de usuarios permite caracterizar formalmente cómo se encuentran conectados los usuarios. Esto es útil para determinar, por ejemplo, cómo se difunde la información de cierto producto entre usuarios, lo que permitiría recomendar nuevos productos a los usuarios o maximizar las ventas de un producto en específico.

Sinha *et al.* [4] mostraron que, si se da a elegir entre las recomendaciones de la red de amigos y las de los usuarios en general, en términos de calidad y utilidad, se prefieren las recomendaciones de los amigos. En consecuencia, los SR que ignoran la estructura de la red social de los usuarios muestran un poder de predicción más limitado.

Por lo anterior, el propósito de esta investigación es evaluar el desempeño del algoritmo de *Matrix Factorization* para la recomendación de películas, teniendo en cuenta modelos de redes sociales inferidas y las medidas topológicas de estas redes.

Descripción del Problema

1.1. Planteamiento del Problema

Actualmente la información en internet crece exponencialmente. Diversas páginas de ventas o servicios contienen diferente material para mostrar a una persona que ingrese a sus sitios web. Los sistemas de recomendación le facilitan a estas empresas a limitar el foco de sugerencias de productos a solo la parte más relevante y de mayor impacto. Estos sistemas son algoritmos que generan recomendaciones a partir de los meta-datos del producto (o del usuario), para ayudar a las personas en procesos de toma de decisiones.

En 2012 Gomez-Rodriguez *et al.* [5] propusieron que detrás de los conjuntos de datos existen enlaces subyacentes desconocidos, sobre los cuales se propaga la información. Al propagarse sobre los enlaces se genera una *cascada* y, gracias a estas, junto con otros datos, se puede inferir redes sociales. El algoritmo desarrollado por Gomez-Rodriguez *et al.* se llama *NETINF*.

Fan *et al.* extendieron *NETINF* al fusionar el trabajo de Gomez-Rodriguez *et al.* junto con datos de calificación de películas, para evaluar el algoritmo de inferencia en función de las recomendaciones [6]. Los autores sugirieron evaluar el algoritmo de Gomez-Rodriguez *et al.* en un conjunto de datos sin red social, como *MovieLens*, para comparar la precisión de las recomendaciones.

En 2017 Kose *et al.* compararon los algoritmos de *User-Based Collaborative Filtering*, *Matrix Factorization* y *Yehuda Koren's Integrated Model*, para predecir las calificaciones que usuarios darían a películas [3]. Esta investigación dio como resultado un desempeño similar para los tres algoritmos, con respecto a su raíz del error cuadrático medio, o *RMSE* (por sus siglas en inglés). Cabe resaltar que se destaca el algoritmo de *Matrix Factorization* gracias a sus buenos resultados y a su sencilla implementación.

En general, los sistemas de recomendación actuales no toman en cuenta las estructuras internas que existen detrás de los conjuntos de datos, las cuales permiten saber cómo se encuentran conectados los usuarios. En consecuencia, surge la pregunta ¿De qué manera se ve afectado el desempeño de los algoritmos de recomendación, haciendo uso de redes sociales inferidas y de las medidas topológicas de estas redes?

1.1.1. Formulación

¿Cómo se ve afectado el desempeño del algoritmo *Matrix Factorization* para la recomendación de películas, al añadir medidas topológicas de redes sociales inferidas?

1.1.2. Sistematización

Se busca responder a las siguientes preguntas:

- ¿Qué algoritmos o métodos existen para inferir redes sociales?
- ¿Qué modelos de redes sociales se pueden inferir a partir de un conjunto de datos en particular?
- ¿Cómo se ve afectada la precisión en las predicciones del algoritmo *Matrix Factorization*, agregando medidas topológicas de redes sociales inferidas?

1.2. Objetivos

1.2.1. Objetivo General

El objetivo de esta tesis es investigar diferentes métodos para la inferencia de redes sociales (con base en el conjunto de datos *MovieLens* [7]), para escoger y analizar un algoritmo, con el fin de evaluar el beneficio de inferir una red social y utilizar las medidas topológicas de esta red, al realizar recomendaciones utilizando el algoritmo *Matrix Factorization*.

1.2.2. Objetivos Específicos

- **O1:** Investigar diferentes métodos o algoritmos para la inferencia de redes sociales.
- **O2:** Explorar los diferentes modelos de redes sociales que se pueden inferir a partir del conjunto de datos de *MovieLens*.
- **O3:** Analizar las métricas de evaluación para el algoritmo *Matrix Factorization*, con el fin de determinar los beneficios de añadir propiedades de redes sociales inferidas.

1.3. Justificación

El centro de esta investigación se fundamenta en encontrar diferentes modelos de redes sociales en el conjunto de datos de *MovieLens*, para evaluar el beneficio de inferir una red social a partir de valoraciones, y utilizar las medidas topológicas de esta red en el algoritmo de recomendación *Matrix Factorization*.

La finalidad de este trabajo es permitir a cualquier organización de servicio usar los algoritmos a desarrollar para mejorar sus sistemas de recomendación. Esto se podría reflejar en recomendaciones para estructuras de comunidad de la red social y no necesariamente a un único usuario.

Mejorar los sistemas de recomendación implica una optimización de recursos, una ampliación de la cobertura y la accesibilidad a sus clientes; lo cual se refleja en mayores beneficios financieros para las organizaciones que adopten estos algoritmos.

1.4. Delimitaciones y Alcances

Teniendo en cuenta los conocimientos en teoría de redes, algoritmia y aprendizaje automático adquiridos en las carreras de Matemáticas Aplicadas e Ingeniería de Sistemas y Computación se pretende extraer diferentes modelos de redes sociales, del conjunto de datos *MovieLens*, para agregar las propiedades de estas redes al algoritmo de recomendación *Matrix Factorization* y comparar las métricas de evaluación de este algoritmo, para determinar los cambios en su desempeño.

De esta forma se alcanzaría el objetivo de evaluar qué tan útil, es inferir una red social a partir de las valoraciones y utilizar las métricas de esta red para realizar recomendaciones de películas. Se prevé que el algoritmo puede ser usado para recomendar no solo películas, sino servicios y productos en general.

1.4.1. Entregables

Archivos ejecutables¹ que contienen el análisis del algoritmo escogido para la inferencia de redes sociales, junto con el cálculo de las medidas topológicas de estas redes, las cuales sirven como insumo para realizar recomendaciones usando el algoritmo *Matrix Factorization*.

¹<https://github.com/suribe06/BSc-Thesis>

Desarrollo del Proyecto

2.1. Marco de Referencia

2.1.1. Áreas Temáticas

A continuación se presentan las categorías relacionadas con este proyecto:

- Information systems → Information systems applications → Data mining → Collaborative filtering.
- Mathematics of computing → Discrete mathematics → Graph theory.
- Networks → Social Networks → Network Properties.
- Computing methodologies → Machine learning → Machine learning approaches → Factorization methods → Non-negative matrix factorization.

2.1.2. Marco Teórico

Para el desarrollo de la presente propuesta, se deben tener en cuenta conocimientos tanto de ciencias de la computación como de matemáticas. A continuación se presentan las bases teóricas sobre las que se trabajará en el desarrollo del proyecto.

2.1.2.1. Factorización de Matrices (FM)

Sea $U = \{u_1, u_2, \dots, u_m\}$ un conjunto de usuarios, $I = \{i_1, i_2, \dots, i_n\}$ un conjunto de ítems. Sea T un conjunto finito dado con:

$$T \subset U \times I \times \mathbb{R}$$

Cada tripleta $(u, i, r) \in T$, con $1 \leq r \leq 5$; significa que el usuario u calificó el ítem i con el valor r . Note que $|T| < m \cdot n$ debido a que la mayoría de usuarios sólo valoran algunos ítems. Ahora, sea R una matriz $m \times n$ que representa a T , donde cada entrada $r_{u,i}$ ($1 \leq u \leq m$ y $1 \leq i \leq n$) tiene el valor de la calificación r del usuario u sobre el ítem i .

Dado un usuario $u \in U$ y un ítem $i \in I$ de quien se desconoce $r_{u,i}$, el objetivo es predecir la valoración de u sobre el ítem i . Para esto, se emplean técnicas de factorización matricial, para aprender las características latentes de los usuarios y los ítems, a fin de predecir las valoraciones

desconocidas, utilizando estas características latentes [8]. La tarea entonces, es descubrir k características latentes para encontrar dos matrices $\mathbf{P} \in \mathbb{R}^{m \times k}$ y $\mathbf{Q} \in \mathbb{R}^{k \times n}$ tales que su producto se aproxime a \mathbf{R} :

$$\mathbf{R} \approx \mathbf{P}\mathbf{Q} = \hat{\mathbf{R}}$$

La matriz \mathbf{P} representa la asociación entre un usuario y las características, mientras que la matriz \mathbf{Q} representa la asociación entre un ítem y las características.

Se puede obtener la predicción $\hat{r}_{u,i}$ de la valoración de un ítem i por parte de un usuario u mediante el cálculo del producto punto de los dos vectores correspondientes a u e i .

$$\hat{r}_{u,i} = p_u q_i = \sum_{d=1}^k p_{ud} \cdot q_{di}$$

Si esto se hace para todos los usuarios e ítems, la matriz de valoración completa usuario-ítem se convierte en:

$$\begin{bmatrix} p_1 \\ \vdots \\ p_m \end{bmatrix} \begin{bmatrix} q_1 & \cdots & q_n \end{bmatrix} = \begin{bmatrix} \hat{r}_{1,1} & \cdots & \hat{r}_{1,n} \\ \vdots & \ddots & \vdots \\ \hat{r}_{m,1} & \cdots & \hat{r}_{m,n} \end{bmatrix} = \hat{\mathbf{R}}$$

Para aprender a reconocer los vectores de características (p_u y q_i), el sistema minimiza la suma del error cuadrático medio ($RMSE$, por sus siglas en inglés), en el conjunto de valoraciones conocidas:

$$\begin{aligned} SSE &= \sum_{(u,i,r) \in T} (r_{u,i} - \hat{r}_{u,i})^2 = \sum_{(u,i,r) \in T} \left(r_{u,i} - \sum_{d=1}^k p_{ud} \cdot q_{di} \right)^2 \\ RMSE &= \sqrt{SSE/|T|} \\ (\mathbf{P}^*, \mathbf{Q}^*) &= \arg \min_{(\mathbf{P}, \mathbf{Q})} RMSE^1 \end{aligned} \tag{1}$$

Koren *et al.* [9] mostraron que estos modelos son muy propensos al sobreajuste si se estiman tal cual, por lo que a menudo se estiman con regularización. Por lo que la ecuación (1) se convierte en:

$$SSE' = \sum_{(u,i,r) \in T} \left(r_{u,i} - \sum_{d=1}^k p_{ud} \cdot q_{di} \right)^2 + \lambda (||q_i||^2 + ||p_u||^2) \tag{2}$$

¹Las matrices \mathbf{P} , \mathbf{Q} para las que el $RMSE$ alcanza su valor mínimo.

$$RMSE' = \sqrt{SSE'/|T|}$$

$$(\mathbf{P}^{*'}, \mathbf{Q}^{*'}) = \arg \min_{(\mathbf{P}, \mathbf{Q})} RMSE'$$

Donde $\|\cdot\|^2$ es la norma de Frobenius y λ es el parámetro de regularización.

Dos enfoques para minimizar la ecuación (2) son el descenso de gradiente estocástico y los mínimos cuadrados alternos (los detalles técnicos se encuentran en [9]).

2.1.2.2. Redes Sociales

Una red N puede definirse como $N = (V, E, f_V, f_E)$, que contiene un grafo $G = (V, E)$, el cual es un par ordenado de un conjunto de vértices V y un conjunto de aristas E , extendido con una función f_V que especifica propiedades de los vértices ($f_V : V \rightarrow X$) y una función f_E que especifica propiedades de las aristas ($f_E : E \rightarrow Y$) [10].

2.1.2.3. Inferencia de Redes Sociales

Gomez-Rodriguez *et al.* [5] presentan cómo inferir una red basándose en un modelo probabilístico generador. Definen una cascada c como una pieza diferente de información que se propaga por una red. Sea \hat{G} una red oculta sobre la que se extienden múltiples cascadas. Para poder inferir esta red se define primero el modelo de transmisión de cascada $P_c(u, v)$, que describe la probabilidad de que el nodo u propague la cascada c , al nodo v .

Luego se describe la probabilidad $P(c|T)$ que la cascada c se propague en un determinado árbol $T = (V_T, E_T)$, donde el árbol de propagación T especifica qué nodos infectaron a otros nodos. Por lo tanto, dado un árbol T , la probabilidad de la cascada c se puede calcular mediante reglas bayesianas así:

$$P(c|T) = \prod_{(u,v) \in E_T} P_c(u, v) \quad (3)$$

Por último, se define $P(c|G)$, como la probabilidad que la cascada c ocurra en una red G . Esta se puede calcular basándose en $P(c|T)$. Se buscan todas las posibles estructuras de arboles T en $\mathcal{T}_c(G)$ en donde la cascada c puede propagarse. Aquí, $\mathcal{T}_c(G)$ es el conjunto de todos los posibles árboles de propagación de la cascada c en la red G . En la ecuación (4) se calcula la probabilidad que una sola cascada c ocurra en una red G .

$$P(c|G) = \sum_{T \in \mathcal{T}_c(G)} P(c|T)P(T|G)$$

$$\propto \sum_{T \in \mathcal{T}_c(G)} \prod_{(u,v) \in E_T} P_c(u, v) \quad (4)$$

Ahora, la probabilidad que un conjunto de cascadas C ocurra en la red G se define como:

$$P(C|G) = \prod_{c \in C} P(c|G) \quad (5)$$

Lo presentado es el modelo de transmisión en cascada. Bajo este modelo, se busca estimar una red \hat{G} que (aproximadamente) maximice la probabilidad que C ocurra en él, es decir, encontrar \hat{G} que resuelva el siguiente problema de optimización.

$$\hat{G} = \arg \max_{|G| \leq e} P(C|G)^2 \quad (6)$$

Note que la maximización es sobre todos las redes dirigidas G de a lo sumo e aristas, y $P(C|G)$ se define mediante la ecuación (5).

Este problema de optimización parece intratable. Para evaluar la ecuación (5), hay que calcular la ecuación (4) para cada cascada $c \in C$ en todos los posibles arboles $T \in \mathcal{T}_c(G)$ y el número de árboles puede ser exponencial, con respecto al tamaño de G . Por esta razón, los autores de [5] introducen dos conceptos para que la ecuación (6) sea tratable. El primer concepto es haciendo uso únicamente del árbol más probable en lugar de todos los árboles de propagación posibles. El segundo concepto lo definen como “ ε -bordes”, el cual hace referencia a que los nodos pueden infectarse por razones diferentes a la influencia de la red.

2.2. Medidas Topológicas de Redes Sociales

Nacher [11] señala que las medidas topológicas de una red se refieren a una variedad de medidas matemáticas, que utilizan la matriz de adyacencia de un grafo G para capturar propiedades específicas de la topología de la red. Algunas de estas medidas son:

- **Grado de un nodo:** Sea un nodo $u \in V$, el grado de u se denota por $\deg(u) = |\{v \in V : \{u, v\} \in E\}|$, e indica el número de aristas conectadas al nodo u . Nodos muy conectados se denominan *hubs*.
- **Distancia:** La distancia $d(u, v)$ entre dos nodos $u, v \in V$ se define como el número más corto de aristas a recorrer, desde el nodo u al nodo v . Si los nodos no son accesibles entonces $d(u, v) = \infty$.
- **Excentricidad:** La excentricidad de un nodo es la distancia de su camino más corto desde el otro nodo más lejano en el grafo. Se denota por $e(u) = \max\{d(u, v) : v \in V\}$
- **Coefficiente de agrupamiento:** Mide la proporción del número de aristas entre los vecinos del nodo u y el número máximo de aristas que podría existir entre los vecinos del nodo u . Sirve para cuantificar qué tanto está agrupado un nodo u con sus vecinos.

²El grafo G para el cual la ecuación (5) alcanza su valor máximo

- **Medidas de centralidad:** Sirven para calcular la importancia de un nodo u en la red.

- Centralidad de Cercanía (o *Closeness Centrality* por su nombre en inglés): Puntúa un nodo u en función de su cercanía a todos los demás nodos de la red. Está dada por:

$$CC(u) = \frac{n}{\sum_{v \in V} d(u, v)}$$

Donde n es el número de nodos del grafo.

- Centralidad de Intermediación (o *Betweenness Centrality* por su nombre en inglés): Cuantifica el número de veces que un nodo se encuentra en el camino más corto entre otros dos nodos. Está dada por:

$$BC(u) = \sum_{u \neq v \neq s \in V} \frac{\sigma_{sv}(u)}{\sigma_{sv}}$$

Donde σ_{sv} es el número total de caminos más cortos, desde el nodo s al nodo v y $\sigma_{sv}(u)$ es el número de esos caminos que pasan por u .

- Centralidad del Vector Propio (o *Eigenvector Centrality* por su nombre en inglés): Esta propiedad mide la influencia de un nodo en una red. Puntúa un nodo u en función de la centralidad de sus vecinos. Está dada por:

$$EC(u) = \frac{1}{\lambda} \sum_{v \in N(u)} EC(v) = \frac{1}{\lambda} \sum_{v \in V} a_{u,v} \cdot EC(v)$$

Donde λ es una constante, $N(u)$ es el conjunto de los vecinos de u y $a_{u,v}$ es una entrada de la matriz de adyacencia A del grafo G .

- Centralidad de Lejanía (o *Farness Centrality* por su nombre en inglés): Esta medida es recíproca a la centralidad de cercanía (de modo que si la cercanía es pequeña, la lejanía es grande y viceversa), es decir, es la suma de la distancia de un nodo a todos los demás nodos del grafo.

Muchas otras métricas que se podrían usar en este proyecto se encuentran definidas en [11].

De igual forma se piensa hacer uso de estructuras de comunidad, que nos permiten revelar relaciones ocultas entre los nodos de la red. Fortunato [12] define las comunidades como grupos de nodos, que probablemente comparten propiedades comunes y/o desempeñan funciones similares dentro del grafo.

2.3. Trabajos Relacionados

En la exploración y búsqueda de información para el desarrollo de este proyecto, se revisó la literatura existente a través de bases de datos, antecedentes de estudios, investigaciones o artículos similares, que permitan construir una base sólida para el desarrollo del presente proyecto de investigación.

- Koren *et al.* [9] presentan diferentes técnicas de factorización matricial para sistemas de recomendación, como la descomposición en valores singulares (SVD por sus siglas en inglés), el descenso de gradiente estocástico y mínimos cuadrados alternos. Debido a su primer puesto en *Netflix Prize Competition* (competencia que buscaba el mejor algoritmo de filtrado colaborativo, para predecir las valoraciones de los usuarios sobre películas) en 2007 y 2008, Koren *et al.* indican que estos son los métodos más exitosos y dominantes dentro de los recomendadores de filtrado colaborativo.
- Ma *et al.* [13] basándose en que la red social de un usuario afectará los comportamientos de dicho usuario, presentan un marco de recomendación social, que fusiona una matriz de calificación de elementos de usuario, con la red social del usuario, utilizando factorización de matriz probabilística.
- He *et al.* [14] proponen un sistema de recomendación basado en redes sociales. Este sistema utiliza la información de las redes sociales, incluidas las preferencias de los usuarios, la aceptación general de los artículos y la influencia de los amigos sociales. A partir de esa información desarrollan un modelo probabilístico, que les permite hacer recomendaciones personalizadas.
- Liu *et al.* [15] plantean que las redes sociales de los usuarios y las opiniones de estos se pueden incorporar para mejorar la precisión de la predicción en el ámbito de la recomendación de películas. Proponen entonces un método de recomendación basado en la red local de confianza (*LTN*). Para esto extraen patrones de redes sociales latentes y las múltiples fuentes de opiniones de los usuarios, para generar la red local de confianza de usuarios y poder realizar la predicción basándose en esta red. El algoritmo *LTN* supera significativamente el rendimiento de la recomendación estándar del Filtrado Colaborativo con respecto al *RMSE* y al *MAE*.
- Wang *et al.* [16] proponen ver el problema de recomendaciones como un problema de predicción de enlaces en redes bipartitas. Para dar solución a esto, propusieron el algoritmo *SRNMF*, que toma explícitamente las características latentes de los nodos, junto con su estructura topológica intrínseca y codifica la información geométrica de la red, mediante la construcción de una matriz basada en similitudes. En comparación con otros 17 métodos de predicción de enlaces, el método *SRNMF* es significativamente superior en términos de precisión y estabilidad.
- Li *et al.* [17] plantean que la factorización matricial (FM) estándar no captura las correlaciones estructurales jerárquicas, por ende, proponen una técnica de Factorización Matricial Jerárquica Oculta (*HHMF*), que aprende la estructura jerárquica oculta a partir de los registros de usuario-artículo. Esta técnica no requiere el conocimiento previo de la estructura jerárquica; por lo tanto, puede aplicarse cuando esta información sea explícita o implícita. *HHMF* supera a los métodos tradicionales de FM, a los métodos de FM jerárquica y a los métodos basados en redes neuronales.

- Gasparetti *et al.* [18] revisan la técnica de recomendación social basada en la detección de comunidades. Hacen énfasis en que grupos de usuarios que tienen características sociales más similares, pueden proporcionar evidencia valiosa y adicional para el proceso de recomendación. Sin embargo, debido a que se requieren varios pasos en las técnicas de detección de comunidades en sistemas de recomendación, esto hace que sea un objetivo principalmente teórico y prácticamente desafiante.

2.4. Métodos para Inferir Redes Sociales

- Gomez-Rodriguez *et al.* [5] propusieron que detrás de ciertos conjuntos de datos existe alguna red estática desconocida subyacente, sobre la cual se propaga la información. Al propagarse sobre la red se genera una *cascada* y gracias a estas, junto con otros datos, se puede inferir redes sociales. Como resultado, basándose en un modelo probabilístico generador, desarrollaron un algoritmo escalable para la inferencia de redes sociales llamado *NETINF*.
- Fan *et al.* [6] plantearon fusionar el trabajo de Gomez-Rodriguez *et al.* junto con datos de calificación de películas. Propusieron el algoritmo *MOVINF* basándose en [5], para predecir una red social mediante cascadas de calificación de películas. El algoritmo propuesto por Fan *et al.* utiliza el concepto de *k* vecinos mas cercanos en el modelo probabilístico generador usado en *NETINF* para calcular la probabilidad de que una cascada se propague entre *k* vecinos mas cercanos. Este algoritmo mejora la complejidad temporal y la precisión del algoritmo *NETINF* bajo ciertas condiciones.
- Alpay *et al.* [19] desarrollan el algoritmo *FASTINF* basado en *NETINF*, para usarlo en un entorno de big data en la industria de grandes redes sociales en línea como Facebook o Twitter. El modelo probabilístico generador usado por Alpay *et al.* es similar al trabajo de [5], pero en lugar de utilizar la estrategia de árbol de propagación mas probable, consideran la cascada como un conjunto de bordes dispersos en un grafo y, por lo tanto, el algoritmo infiere arcos en lugar del árbol. El algoritmo desarrollado obtiene ordenes de complejidad temporal mas rápidos sin sacrificar la precisión.
- Gomez-Rodriguez *et al.* [20] proponen un algoritmo paralelo denominado *NETRATE*, el cual se basa en su algoritmo *NETINF*. La diferencia radica que en [5] la velocidad de transmisión es fija, y no inferida. Este nuevo algoritmo se basa entonces en encontrar la red óptima y las velocidades de transmisión que maximizan la probabilidad de un conjunto observado de cascadas. La principal novedad de este método es modelar la difusión como una red de sucesos temporales continuos e independientes que ocurren a diferentes velocidades.
- Gomez-Rodriguez *et al.* [21] proponen un nuevo algoritmo el cual extiende su algoritmo previo *NETINF*, la principal innovación de este algoritmo es abordar el problema de la inferencia de red considerando todos los árboles posibles admitidos por un grupo de cascadas. Los autores presentan un algoritmo de inferencia de red que puede ser capaz de inferir redes del orden de cientos de miles de nodos con una pequeña cantidad de cascadas observadas. Este algoritmo

no infiere probabilidades previas de infección ni tasas de transmisión, sino solo la conectividad de la red.

- Gray *et al.* [22] utilizan el modelo de cascada independiente (aunque se puede extender a cualquier modelo en cascada) para desarrollar un método de inferencia bayesiano basado en la Cadena de Markov Monte Carlo, teniendo en cuenta las incertidumbres inherentes que surgen en la inferencia y las observaciones de datos. Los autores utilizan el modelo de inferencia bayesiano junto al “Teorema del Árbol de Matrices” de Kirchoff, extendido a árboles dirigidos por Tutte [23], en el proceso de la inferencia de la red. Esta técnica es probada con un pequeño número de cascadas simuladas para las cuales *NETINF* no produce un resultado.

2.5. Metodología

2.5.1. Tipo de Estudio

Esta investigación es de tipo teórico-experimental, pues es necesario crear un modelo computacional, el cual incluye la inferencia de redes sociales, las métricas de estas redes y el algoritmo *Matrix Factorization* para la recomendación. Una vez el modelo está planteado, este se prueba con el conjunto de datos *MovieLens*.

2.5.2. Actividades

Para alcanzar los objetivos específicos de esta investigación, se realizarán las siguientes actividades:

- **A1: Análisis de los datos y almacenamiento en la base de datos.** Hacer un primer acercamiento al conjunto de datos *MovieLens*. Posteriormente, almacenar esta información en Neo4j para poder visualizarla.
- **A2: Revisión sistemática de la literatura.** Revisar los estudios previos que se reportan en artículos, libros, conferencias, etc. relacionados con sistemas de recomendación e inferencia de redes sociales.
- **A3: Investigación de Métodos.** Investigación de diversos métodos para la inferencia de redes sociales. Y elección del algoritmo.
- **A4: Exploración de modelos en redes sociales.** A partir del algoritmo escogido, explorar los diferentes modelos de redes sociales que se pueden generar.
- **A5: Elección de métricas a usar.** Determinar las mejores métricas de redes sociales para realizar predicciones.
- **A6: Extracción de métricas a las redes sociales.** Hacer uso de algoritmos para la extracción de las métricas escogidas.

- **A7: Evaluación del trabajo propuesto.** Probar el modelo obtenido con el conjunto de datos *MovieLens*, con el fin de analizar las métricas de evaluación (*accuracy*, *precision*, *recall*, *f1-score*) del algoritmo *Iterative Matrix Factorization*.
- **A8: Documentación.** Presentación de los resultados obtenidos y conclusiones.

2.6. Resultados Esperados

- Evidenciar la aplicación de los conocimientos adquiridos en las carreras de Matemáticas Aplicadas e Ingeniería de Sistemas y Computación.
- Obtener buenos resultados en cuanto a las predicciones, evidenciándose en las métricas de evaluación de *Matrix Factorization*.
- Entender como el uso de las redes sociales junto con sus propiedades, podría mejorar potencialmente la precisión de los sistemas de recomendación.
- Contribuir positivamente en el ámbito de los sistemas de recomendación y la inferencia de redes sociales.

2.7. Cronograma

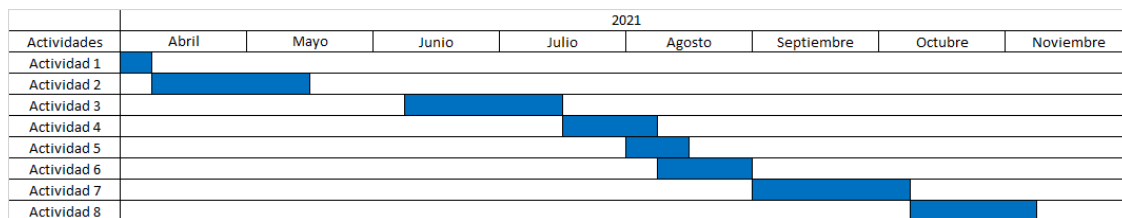


Figura 2.1: Diagrama de Gantt del cronograma de actividades

2.8. Recursos

2.8.1. Recursos Humanos

- Jorge Finke Ortiz, docente del Departamento de Ingeniería Eléctrica y Ciencias de la Computación, miembro del grupo de investigación GAR, y miembro activo en la organización IEEE, es el director de esta propuesta.
- Carlos Ernesto Ramírez Ovalle, docente del Departamento de Ciencias Naturales y Matemáticas y miembro del grupo de investigación EMAP, es el codirector de esta propuesta.

- Santiago Uribe Pastás, estudiante de Ingeniería de Sistemas y Computación - Matemáticas Aplicadas, es el responsable de realizar la investigación propuesta.

2.8.2. Recursos Técnicos

- Lenguaje de programación Python (junto con algunas de sus librerías), para construir algoritmos que transformen, procesen y analicen los datos correspondientes.
- Suscripciones a los diferentes motores de búsqueda académicos (IEEE, ACM, Scopus y SpringerLink), para adquirir los artículos completos necesarios para esta investigación.
- Base de datos orientada a grafos Neo4j, para poder cargar y visualizar los datos de la manera necesaria, para el enfoque propuesto.
- Computador ASUS X441UV con un procesador Intel Core i5-7200U y 16 GB de RAM.

2.8.3. Presupuesto

En la Tabla 2.1 se presentan los valores de todos los recursos a usar en el proyecto.

Recurso	Cantidad	Valor
Recursos Humanos	3	12,000,000
Software	3	0
Computador	1	2,500,000
Total	-	14,500,000

Tabla 2.1: Tabla de Presupuesto

Bibliografía

- [1] Y. Fang and L. Si, “Matrix Co-factorization for Recommendation with Rich Side Information and Implicit Feedback,” in *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems - HetRec '11*, (New York, New York, USA), pp. 65–69, ACM Press, oct 2011.
- [2] Netflix, “How Netflix’s Recommendations System Works.” <https://help.netflix.com/en/node/100639/us>. Online. [Accessed: 27- Mar- 2021].
- [3] A. Kose, C. Kanbak, and N. Evirgen, “Performance comparison of algorithms for movie rating estimation,” in *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, pp. 955–959, Institute of Electrical and Electronics Engineers Inc., 2017.
- [4] R. R. Sinha and K. Swearingen, “Comparing Recommendations Made by Online Systems and Friends,” in *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, (Dublin, Ireland), jun 2001.
- [5] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, “Inferring networks of diffusion and influence,” in *ACM Transactions on Knowledge Discovery from Data*, vol. 5, (New York, New York, USA), pp. 1–37, ACM PUB27, feb 2012.
- [6] C. Fan and L. Yu, “Inferring Social Networks Based on Movie Rating Data,” tech. rep., Stanford University, CA, 2011.
- [7] F. M. Harper and J. A. Konstan, “The movielens datasets: History and context,” *ACM Transactions on Interactive Intelligent Systems*, vol. 5, pp. 1–19, dec 2015.
- [8] G. Takács, I. Pilászy, B. Németh, and D. Tikk, “Matrix factorization and neighbor based algorithms for the netflix prize problem,” in *RecSys’08: Proceedings of the 2008 ACM Conference on Recommender Systems*, (New York, New York, USA), pp. 267–274, ACM Press, 2008.
- [9] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” *Computer*, vol. 42, no. 8, pp. 30–37, 2009.
- [10] W. de Nooy, “Social Network Analysis, Graph Theoretical Approaches to,” in *Encyclopedia of Complexity and Systems Science*, pp. 8232–8233, Springer New York, jul 2009.
- [11] J. C. Nacher, “Network Metrics,” in *Encyclopedia of Systems Biology*, pp. 1516–1517, Springer New York, jun 2013.
- [12] S. Fortunato, “Community detection in graphs,” *Physics Reports*, vol. 486, pp. 75–174, jun 2009.

- [13] H. Ma, H. Yang, M. R. Lyu, and I. King, “SoRec: Social recommendation using probabilistic matrix factorization,” in *International Conference on Information and Knowledge Management, Proceedings*, (New York, New York, USA), pp. 931–940, ACM Press, 2008.
- [14] J. He and W. Chu, “A social network-based recommender system (snrs),” in *Data Mining for Social Network Data*, vol. 12, pp. 47–74, jun 2010.
- [15] B. Liu and Z. Yuan, “Incorporating Social Networks and User Opinions for Collaborative Recommendation: Local Trust Network based Method,” in *Proceedings of the Workshop on Context-Aware Movie Recommendation - CAMRa '10*, (New York, New York, USA), ACM Press, 2010.
- [16] W. Wang, X. Chen, P. Jiao, and D. Jin, “Similarity-based Regularized Latent Feature Model for Link Prediction in Bipartite Networks,” *Scientific Reports*, vol. 7, dec 2017.
- [17] H. Li, Y. Liu, Y. Qian, N. Mamoulis, W. Tu, and D. W. Cheung, “HHMF: hidden hierarchical matrix factorization for recommender systems,” *Data Mining and Knowledge Discovery*, vol. 33, pp. 1548–1582, nov 2019.
- [18] F. Gasparetti, G. Sansonetti, and A. Micarelli, “Community detection in social recommender systems: a survey,” *Applied Intelligence*, pp. 1–21, nov 2020.
- [19] A. Alpay, D. Demir, and J. Yang, “FastInf: A Fast Algorithm to Infer Social Networks from Cascades,” tech. rep., Stanford University, CA, 2011.
- [20] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf, “Uncovering the temporal dynamics of diffusion networks,” in *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11*, (Madison, WI, USA), p. 561–568, Omnipress, 2011.
- [21] M. Gomez-Rodriguez and B. Schölkopf, “Submodular inference of diffusion networks from multiple trees,” *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, vol. 1, pp. 489–496, may 2012.
- [22] C. Gray, L. Mitchell, and M. Roughan, “Bayesian inference of network structure from information cascades,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 6, pp. 371–381, apr 2020.
- [23] W. T. Tutte, “The dissection of equilateral triangles into equilateral triangles,” *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 44, no. 4, p. 463–482, 1948.