Prototipo de un sistema de recomendación de libros para la biblioteca Mario Carvajal

Stefania Carvajal Ramírez

Universidad del Valle
Facultad de Ingeniería
Escuela de Ingeniería de Sistemas y Computación
Tuluá
2018

Prototipo de un sistema de recomendación de libros para la biblioteca Mario Carvajal

Stefania Carvajal Ramírez
Código 1356165
stefania.carvajal@correounivalle.edu.co

Documento presentado como requisito parcial para la obtención de grado de Ingeniero de Sistemas

Director
Joshua David Triana Madrid, Ing
joshua.triana@correounivalle.edu.co

Universidad del Valle
Facultad de Ingeniería
Escuela de Ingeniería de Sistemas y Computación
Tuluá
2018

Trabajo de grado presentado por
Stefania Carvajal Ramírez
Como requisito parcial para la obtención del título de Ingeniero de Sistemas

Joshua David Triana Madrid
Director

Jurado

Jurado

A mi familia por el apoyo incondicional y constante
A mi director Joshua Triana por ser el mejor por siempre

Tabla de Contenido

Lista de Figuras

Lista de tablas

Resumen

En este proyecto de trabajo de grado se diseña y desarrolla una aplicación prototipo de recomendaciones de libros para la biblioteca Mario Carvajal.

La información con la que se trabaja durante todo el desarrollo fue proporcionada por la biblioteca Mario Carvajal, además se incluye una base de datos de libros externa para realizar pruebas con variables que no se encontraban en la base de datos inicial.

El prototipo web integra dos técnicas de recomendación como lo son el filtrado colaborativo y las recomendaciones basadas en contenido. Con los resultados obtenidos en las pruebas de estos métodos, se realiza una comparación y se define cuál técnica se considera más acorde para una posible integración del prototipo con el sistema de la biblioteca Mario Carvajal.

Capítulo 1

Introducción

1.1. Introducción

La Inteligencia Artificial (IA) es un área de las ciencias de la computación que busca explicar y emular el comportamiento inteligente en términos de procesos computacionales [1]. Entre los muchas técnicas asociadas a la IA se encuentran las redes neuronales, la minería de datos, el procesamiento del lenguaje natural, sistemas distribuidos con agentes inteligentes, la robótica, el aprendizaje de máquinas y quizá la más ambiciosa, que combina muchas de estas ramas, la réplica del cerebro humano.

El aprendizaje de máquinas (Machine Learning) busca automatizar el proceso de aprendizaje de una computadora para que pueda solucionar problemas complejos o computacionalmente costosos [2]. El aprendizaje de máquinas cuenta con múltiples técnicas que se ajustan y responden de mejor manera de acuerdo al tipo de problema al que se aplique. Entre los problemas estudiados están el análisis y detección de patrones en un conjunto de datos, análisis y reconocimiento de imágenes o voz, análisis en la secuencia de ADN, optimización de algoritmos, entre otros.

Dentro del análisis y detección de patrones en un conjunto de datos se incluyen los sistemas de recomendación que analizan, clasifican y filtran colecciones de datos similares y relacionados a un modelo determinado. Los sistemas de recomendación son utilizados frecuentemente por empresas u organizaciones que buscan lucro o mejoras en la prestación de servicios a sus clientes, siendo los más populares aquellos que clasifican películas, música, noticias, libros, artículos, búsquedas, parejas, o cualquier elemento relacionado a la vida y el disfrute social.

Aunque los sistemas de recomendación son más notables en servicios que se prestan en la red (por ejemplo Youtube, Amazon, Spotify, Netflix, entre otros) y que buscan lucro, existen organizaciones públicas sin ánimo de lucro que también han empezado a adoptar estos sistemas. Las bibliotecas públicas, sean físicas o virtuales, han empezado a implementar algoritmos inteligentes que mejoran los resultados de las consultas que realizan sus usuarios. Así pretenden atraer nuevos usuarios que encuentran ventajas en el uso de herramientas tecnológicos, o mantener a los usuarios frecuentes que encontrarán mejoras en el servicio que las bibliotecas pueden ofrecer. Es esto precisamente lo que se quieren lograr en la biblioteca Mario Carvajal de la Universidad del Valle, donde frecuentan todo tipo de usuarios, aquellos que nacieron y crecen en un mundo donde todo está relacionado con la tecnología, y aquellos a los que se les dificulta el uso de las herramientas tecnológicas pero que están dispuestos a aprovechar sus beneficios.

1.2. Objetivos

1.2.1. Objetivo general

Desarrollar un prototipo funcional de recomendación de libros para la biblioteca Mario Carvajal de la Universidad del Valle sede Cali, en una plataforma web utilizando una técnica de Machine Learning.

1.2.2. Objetivos específicos

1. Realizar el preprocesamiento de la información de las bases de datos suministradas por la biblioteca Mario Carvajal aplicando reducciones verticales y horizontales en cada una de las tablas.
2. Seleccionar dos técnicas de Machine Learning para la solución del problema en la biblioteca.
3. Implementar las estrategias seleccionadas con la información que fue previamente procesada.
4. Realizar pruebas con cada técnica utilizando la información procesada de las bases de datos de la biblioteca Mario Carvajal.
5. Analizar los resultados obtenidos para definir la mejor técnica de Machine Learning que servirá de apoyo para el sistema de recomendación de libros en un ambiente bibliotecario.

1.3. Resultados esperados

1.3.1. Relación entre los objetivos específicos y los resultados esperados

Tabla 1.1: Objetivos específicos y su correspondiente resultado

Objetivo específico	Resultados esperados
Realizar el pre-procesamiento de la información de las bases de datos suministradas por la biblioteca Mario Carvajal aplicando reducciones verticales y horizontales en cada una de las tablas.	Base de datos con los atributos y tuplas necesarios y relevantes para la implementación de las técnicas. (Se eliminará cualquier registro que afecte la privacidad o seguridad de los usuarios)
Seleccionar dos técnicas de Machine Learning para la solución del problema en la biblioteca.	Informe con una tabla comparativa de las técnicas de recomendación más utilizadas en el área de bibliotecas y la correspondiente justificación de la elección

Tabla 1.1: Objetivos específicos y su correspondiente resultado

Objetivo específico	Resultados esperados
Implementar las estrategias seleccionadas con la información que fue previamente procesada.	▪ Documento con la definición de los modelos matemáticos de cada técnica aplicados a la situación de la Biblioteca Mario Carvajal. ▪ Documentación relacionada al desarrollo del prototipo de software (requerimientos, diagramas, mockups) ▪ Código de la implementación del prototipo del sistema web
Realizar pruebas con cada técnica utilizando la información procesada de las bases de datos de la biblioteca Mario Carvajal.	▪ Definición de las variables a utilizar en las pruebas. ▪ Explicación de los resultados que se esperan obtener con cada una de las técnicas.
Analizar los resultados obtenidos para definir la mejor técnica de Machine Learning que servirá de apoyo para el sistema de recomendación de libros en un ambiente bibliotecario.	Informe con el análisis estadístico de los resultados obtenidos con cada técnica al realizar las pruebas.

1.4. Planteamiento del Problema.

Siendo la Universidad del Valle una de las mejores universidades a nivel regional [3], procura igualar su oferta académica y sus servicios con aquellas universidades que tienen un modelo más destacado a nivel nacional. Es reconocida gracias a las diferentes sedes que se encuentran distribuidas por todo el departamento del Valle del Cauca. Una de las razones que la hacen merecedora de tal reconocimiento es la Biblioteca "Mario Carvajal" ubicada en su sede principal. Se encuentra abierta a los estudiantes de todas las sedes de Univalle y al público en general. Actualmente tiene disponibles más de 900.000 ejemplares de todo tipo y en el último año contó con más de 600.000 visitas y más de 150.000 préstamos internos y externos [4].

Cada semestre, la Universidad del Valle programa una inducción a los nuevos estudiantes para que puedan acceder a los diferentes servicios e instalaciones de la división de bibliotecas (que se encuentra conformada por nueve bibliotecas de las sedes regionales y tres bibliotecas en Cali). La biblioteca Mario Carvajal cuenta con más de 20 bibliotecarios para la colección general, cuenta con colección de reserva, mediateca, hemeroteca, exposiciones y servicios para los usuarios con alguna discapacidad.

El sistema bibliotecario tiene diferentes alternativas de consulta de su material disponible en físico y del material al que se puede acceder en línea. Cuenta con una plataforma online de consulta, en donde se puede ingresar el título, autor, materia o alguna palabra clave del tema que se desea consultar en los libros, de esta manera el sistema arrojará como resultado solamente los libros en los cuales el dato ingresado se

encuentra textualmente en alguno de sus atributos en la base de datos. Otra alternativa de consulta es visitar las instalaciones físicas de cualquier biblioteca y explorar los estantes en busca de textos que llenen las necesidades o intereses de los usuarios.

Considerando el gran flujo de personas accediendo a los servicios y el catálogo disponible en comparación con el número de personal bibliotecario en la sede principal, se observa cómo las alternativas de consulta pueden presentar dificultades para algunos usuarios. Para ejemplificar los problemas se definirán dos tipos de usuario: primero se tienen aquellos usuarios con un tema de consulta específico y definido, que saben cuáles textos o materiales cumplirán sus expectativas de consulta, que han realizado préstamos anteriormente y conocen el material que les es más útil. Por otro lado, están los usuarios inexpertos, que poco visitan o hacen uso de los servicios bibliotecarios, que quieren explorar un tema genérico o un género literario, aquellos que buscan consejos o sugerencias de alguien con más experiencia en sus consultas.

El tipo de usuario con más probabilidad de no satisfacer su necesidad de consulta será el inexperto, esto por la manera en que la plataforma online despliega los resultados. Si un usuario no está seguro de qué quiere o necesita consultar, no podrá realizar las consultas con palabras claves o específicas eficientemente y por lo tanto el sistema no será de gran ayuda. Ahora, podría dirigirse hasta las instalaciones físicas de la biblioteca, explorar los estantes en busca de un texto interesante a simple vista, pero esta actividad se convierte en una opción muy poco práctica al tener en cuenta el número de libros disponibles al público o la ubicación y las secciones en que se encuentra el material organizado. La actividad de exploración podría volverse completamente abrumadora y desgastante, por lo que también se considera una alternativa inviable para un usuario inexperto. La siguiente alternativa disponible para este tipo de usuarios será recurrir a la asesoría de un bibliotecario.

La biblioteca principal cuenta con distintas colecciones que necesitan estar constantemente organizadas y monitoreadas por bibliotecarios que son los funcionarios más expertos en el área. Dentro de las muchas funciones de estos profesionales, se encuentra el asesoramiento a los usuarios de la biblioteca, haciendo más fácil el acceso a las diferentes fuentes de información y guiando sobre su uso [5]. Muchas veces los bibliotecarios realizan recomendaciones y sugerencias frente a las peticiones de los usuarios, que dependiendo del tema, la experiencia o la necesidad del usuario, pueden ser acertadas en algunas ocasiones. Estas recomendaciones se hacen basados en la experiencia, sea la personal del bibliotecario o la que ha adquirido frente al tiempo en que lleva ejerciendo su labor. Aún así, por la cantidad de información (ya sea de libros, de usuarios o de las transacciones entre estos) que se maneja en la biblioteca, resulta complicado para un bibliotecario mantener un seguimiento personal con cada usuario, que sería lo más óptimo si se quiere brindar un buen servicio al público. Con las transacciones que realizan los usuarios se tiene una gran fuente de información que reposa almacenada en la base de datos de la biblioteca pero que está siendo desperdiciada al no aplicarse ningún tipo de procesamiento de datos.

Para mejorar el servicio a los usuarios, la calidad de los resultados que arrojan las consultas hechas por éstos y aprovechar la información recolectada en las transacciones, se propone implementar un sistema de recomendación de libros. En los últimos años son muchas las organizaciones que han desarrollado este tipo de sistemas para mejorar sus servicios, haciendo uso de una de las aplicaciones más modernas de la inteligencia artificial: Machine Learning. Teniendo en cuenta los diferentes métodos que se encuentran en el estado del arte para construir un sistema de recomendación, se pretende seleccionar e implementar los mejores antecedentes para realizar un análisis comparativo que logre determinar la mejor estrategia de apoyo para la recomendación de libros en un sistema bibliotecario.

1.5. Antecedentes

Se presentan algunos de los trabajos más representativos respecto a los avances logrados y las técnicas utilizadas para desarrollar sistemas de recomendación de libros.

Comparación de las técnicas filtrado colaborativo, minería por reglas de asociación y algoritmo de Amazon [6]

Los autores realizaron pruebas usando 3 técnicas de recomendación: filtrado colaborativo, reglas de asociación y el algoritmo implementado por Amazon. Pretendían encontrar el mejor método para la recomendación de libros y los mejores atributos o variables para ejecutar las técnicas. Para el filtrado colaborativo hicieron uso de la siguiente información: ID del usuario, ID del material que prestó cada usuario, información bibliográfica y la fecha de devolución del libro. Para las otras técnicas pidieron a un número de usuarios (con los que realizaron las pruebas) el nombre de un libro que quisieran prestar en el momento. Después de ejecutar cada uno de los métodos, les mostraron los resultados a los usuarios seleccionados y les pidieron que calificaran su nivel de interés. La conclusión encontrada es que el filtrado colaborativo puede arriesgar la privacidad de los usuarios y es más costoso computacionalmente, los mejores resultados fueron obtenidos con el algoritmo de Amazon y reglas de asociación.

Combinación de variables para la recomendación y comparación de métodos de ML con el usado por Amazon [7]

Combinaron diferentes variables para implementar las técnicas: los registros de préstamos, título de los libros, categorías NDC (sistema de clasificación bibliotecario chino), año de publicación y el número de veces que se ha prestado un libro, también le pidieron a un grupo de usuarios un libro en el que estarían interesados en prestar. Usaron 3 técnicas: Support vector machine, Adaboost y Rainforest. Recomendaron libros usando las 3 técnicas y cuatro combinaciones de parámetros, también usando similitudes entre el NDC del libro que querían prestar los usuarios y los del catálogo. Finalmente recomendaron los que se obtenían con el algoritmo de Amazon y se encontraban en la biblioteca. Pidieron a los usuarios calificar los resultados y concluyeron que usar el método SVM utilizando los parámetros de los préstamos de los libros, títulos de libros, categorías NDC y el número de veces que el libro fue prestado, tiene más precisión incluso que el algoritmo de Amazon.

Transformar el tiempo de préstamo de un libro en una medida de calificación, filtrado colaborativo híbrido [8]

Hicieron pruebas con dos técnicas de filtrado colaborativo: k-vecinos cercanos y latent factor model. Utilizaron diferentes variables para realizar varias pruebas individuales con cada algoritmo. Después utilizaron dos modelos para mezclar las técnicas: regresión lineal y redes neuronales. Así, hicieron pruebas mezclando vecinos cercanos de elementos con vecinos cercanos de usuarios y vecinos cercanos de usuarios con factor latente, cada una con regresión lineal y con redes neuronales. Los resultados finales mostraron que es mejor usar la mezcla de las técnicas que implementarlas individualmente, además que se obtienen predicciones más precisas al hacer uso de redes neuronales.

Recomendaciones más personalizadas usando datos demográficos y web scraping para el problema de "cold start" [9]

Los autores querían crear confianza en el sistema haciendo las recomendaciones más personalizadas y explicando el funcionamiento de su sistema a los usuarios. Usaron 3 técnicas de recomendación: filtrado colaborativo, basado en contenido y datos demográficos. Los datos demográficos son usados para filtrar los resultados obtenidos con las otras dos técnicas y crear clusters dependiendo de la variable demográfica seleccionada. Pretenden aliviar el problema de cold start usando "web scraping" desde portales populares

para dar recomendaciones trending.

Solucionar problema de escasez de datos y cold start combinando CF y reglas de asociación [10]

Utilizan el filtrado colaborativo basado en la similitud entre elementos para realizar las recomendaciones, para calcular esta similitud usan adjusted cosine vector con el rating que cada usuario asigna a los elementos. Para manejar el problema de falta de datos (pues no todos los elementos han sido calificados) usan reglas de asociación. Luego realizan las recomendaciones usando la estrategia de vecinos más cercanos. Comparan sus resultados con los que daría la base de datos mostrando resultados filtrados por autor. También encuentran mejores resultados con un número de "vecinos" más alto.

"Cascada" de filtrado basado en contenido y colaborativo [11]

Realizan la combinación de los dos métodos en forma de cascada: primero extraen la información necesaria o relevante de la base de datos, después dividen o hacen clusters de los libros en géneros y estos en subgéneros, así le aplican el filtrado basado en contenido de acuerdo a lo ingresado por el usuario, con estos resultados aplican filtrado colaborativo basado en el rating de los libros elegidos y los que ha calificado el usuario previamente y finalmente se muestran los resultados obtenidos para ser recomendados. No analizan sus resultados ni realizan comparaciones pero muestran el aspecto del sistema web y los módulos que implementan.

Recomendación de clusters de libros, influencia de carreras, materias en préstamos [12]

Emplean la técnica de filtrado colaborativo basado en elementos, usan los registros de préstamos pero proponen una nueva técnica basada en probabilidad pues aseguran que la carrera y el año de estudio de los usuarios influye en los préstamos que realiza y en sus intereses. Las recomendaciones son realizadas en forma de clusters de libros que pertenecen a una misma categoría y no por libros individuales, pues utilizan las categorías del método de clasificación chino usado en la biblioteca elegida y la frecuencia en que un usuario realiza transacciones por cada categoría como valor de calificación. Los resultados obtenidos con el algoritmo que propusieron fueron mucho más precisos que con la técnica de filtrado colaborativo regular, pero esta última arrojó resultados más novedosos que tal vez el usuario nunca hubiese considerado por sí solo. Mencionan que tal vez un híbrido entre las dos técnicas resulte más ventajoso en términos de precisión.

Calificación inicial para los libros nuevos [13]

Implementan una calificación inicial de acuerdo al resumen, autor y prefacio de los libros nuevos, asignada por un equipo de catalogación o un bibliotecario experto. Tienen en cuenta factores que afectan la calificación de algunos usuarios como el tipo de usuario (si es profesor o profesor asociado tiene más influencia), la cantidad de libros leídos en una misma categoría (más libros leídos, mayor experiencia para calificar la categoría) y si la calificación dada por un usuario está cercana a la calificación promedio del libro. Estos factores influyen en el cálculo de similitud entre usuarios. Los lectores se categorizan en máximo 3 tipos de categorías a la vez. No se presentan análisis de los resultados, se enfocan más en mostrar los módulos definidos para construir el sistema de recomendación.

1.6. Justificación

1.6.1. Justificación metodológica

Con la afluencia de usuarios que llegan diariamente a la biblioteca Mario Carvajal, sea para realizar consultas internas o externas, se están generando transacciones en el sistema bibliotecario que conforman una gran fuente de información disponible para ser explotada. Con el crecimiento exponencial de la red y los avances alcanzados en los últimos años en materia de Inteligencia Artificial, la minería de datos empieza a florecer para aprovechar la información que se genera en cada momento cuando todos navegamos en internet. La recolección de información y el procesamiento de la misma ahora es mucho más frecuente entre todas las organizaciones que ofrecen productos y servicios, incluso las públicas como las bibliotecas. Las diferentes técnicas de IA que procesan datos y encuentran patrones o filtran información son especialmente útiles para hacerle un seguimiento mucho más personal a cada usuario. De esta manera, aprovechando la información que se obtiene con cada consulta y transacción que se hace en la biblioteca Mario Carvajal y considerando el número tan grande de usuarios con los que cuenta y el poco personal disponible en comparación, implementar una técnica de IA para la recomendación de libros es un camino viable. Se eligen técnicas de Machine Learning para ser analizadas y aplicadas en el problema considerado por los antecedentes que se encuentran respecto a trabajos realizados en bibliotecas físicas o virtuales, y los buenos resultados que se han obtenido. [8, 6, 7, 10, 12]

1.6.2. Justificación práctica

El sistema de recomendación clasificaría y filtraría los textos con similitudes en su contenido, ofreciendo más alternativas de consulta a los usuarios y realizando sugerencias más personalizadas que se basarán en los préstamos realizados por ellos anteriormente. De esta manera, la cantidad de bibliotecarios que deberían estar disponibles para las consultas de los usuarios no tendría que ser tan alta, además podrían aumentarse las transacciones que se hacen en el sistema bibliotecario si las recomendaciones tienen un nivel de certeza mayor al promedio respecto a las pruebas realizadas.

1.7. Información sobre los capítulos

En la siguiente tabla se muestra una corta descripción de cada capítulo del documento.

Tabla 1.2: Estructura de capítulos

Capítulo	**Descripción**	**Objetivo específico**
Capítulo 2: Marco referencial	Se muestra el estado del arte referente a las aplicaciones de los sistemas de recomendación en temas generales, así como también algunos ejemplos de trabajos realizados en sistemas de bibliotecas	Objetivo 2
Capítulo 3: Alcances de la propuesta	En este capítulo se definen los alcances de la propuesta	Ninguno
Capítulo 4: Modelos	Se presentan los modelos matemáticos para implementar los algoritmos de recomendacion.	Objetivo 3
Capítulo 5: Metodología	En este capítulo se describe la metodología utilizada durante el desarrollo de la aplicación prototipo.	Objetivo 3
Capítulo 5: Desarrollo	En este capítulo se describe el proceso de desarrollo del proyecto.	Objetivo 3

Capítulo 6: Pruebas	El capítulo muestra la descripción de las pruebas y los resultados obtenidos en éstas.	Objetivo 4 y 5
Capítulo 7: Conclusiones	Se enumeran las conclusiones definidas al terminar el proyecto.	Ninguno.
Capítulo 8: Trabajos futuros	Se describen mejoras que pueden implementarse a futuro.	Ninguno

Capítulo 2

Marco referencial

2.1. Introducción

Con el crecimiento exponencial de internet, se empezó a evidenciar el aumento considerable de la información que se acumula en la web y se presenta a los usuarios[14]. Con tantos sitios y opciones para consultar, internet pasa de ser una gran herramienta a convertirse en un problema para el usuario que puede sentir una sobrecarga de información incapaz de procesar o manejar adecuadamente. Una de las soluciones que se están implementado para enfrentar este problema son los sistemas de recomendación.[14]

El propósito principal detrás de estos algoritmos es reducir el ruido en la información y filtrar aquella que es realmente relevante para las compañías y los usuarios[15, 14]. Los sistemas de recomendación se encargan de mostrar sugerencias o recomendaciones personalizadas con base en los gustos y necesidades de los usuarios, mejorando la experiencia que éstos tienen en el sitio web e influenciando la construcción de la confianza y lealtad del usuario respecto a la compañía. Los algoritmos de recomendación surgieron principalmente como una ayuda para las organizaciones de comercio electrónico[16, 17]. En la mayoría de compañías que ofrecen productos o servicios, el objetivo principal es suplir las necesidades del cliente o usuario final para obtener un beneficio económico y es precisamente esto lo que se logra al implementar un sistema de recomendación, pues ayuda a predecir las necesidades del usuario con la mayor precisión posible.

Estos sistemas deben contar con ciertas características o funcionalidades que son importantes para que las recomendaciones sean satisfactorias para los usuarios. La calidad de las recomendaciones se nota en la relevancia que éstas tienen respecto a los intereses del usuario. Sin embargo, el usuario también quiere contar con cierto grado de novedad, que le permita explorar nuevos intereses para que pueda ampliar el espectro de sus gustos o necesidades. Existe un factor que se menciona frecuentemente cuando se hace referencia a los sistemas de recomendación y es la serendipia, que aunque es muy similar a la novedad (pues permite encontrar nuevos productos relacionados a los gustos establecidos) cuenta con un grado de sorpresa, la recomendación es "inesperada" pero logra captar la atención del usuario. Otro factor importante es la seguridad y privacidad de la información que se maneja de los usuarios, es éste tal vez la característica más importante a la hora de influenciar el nivel de confianza que los usuarios le darán al sistema.[16]

Además de las características mencionadas anteriormente, para que un sistema de recomendación sea realmente útil, debe contar con una gran cantidad de datos que puedan ser procesados. Los sitios web los recopilan implícitamente cuando capturan los movimientos o interacciones que tienen los usuarios con la página web, o explícitamente cuando le piden información directamente al usuario sobre sus gustos e

intereses[16]. Así, logran crear un perfil de usuario que luego puede relacionarse con perfiles de otros usuarios o con los productos ofrecidos para efectuar las recomendaciones. Haciendo uso de alguna técnica de machine learning o minería de datos, se encuentran patrones o correlaciones que dan razón sobre posibles productos (usualmente llamados ítems) que pueden ser de interés para el usuario y que éste aún no conoce.

Actualmente los sistemas de recomendación se ven principalmente en las compañías que ofrecen contenidos a los usuarios como noticias, música, películas, libros, o productos en general como es el caso de Amazon (quien fue uno de los pioneros junto con GroupLens, en la investigación y desarrollo de los sistemas de recomendación[16]). Sin embargo por la gran utilidad que representan estos algoritmos, se están presentando aplicaciones en compañías que prestan servicios como por ejemplo turismo, transporte, salud y educación.

Los sistemas de recomendación se basan en la premisa de que siempre existen dependencias significativas entre las dimensiones de los usuarios versus ítems[16], esta correlación se puede explotar haciendo uso de alguno de los métodos que analizan la información de estas dos entidades. Es frecuente encontrar trabajos en los que se utiliza un enfoque basado en el contenido de la información, también se encuentran trabajos que hacen uso del filtrado colaborativo y algunos optan por proponer un algoritmo híbrido que aproveche las ventajas de cada uno de los métodos anteriores.

Sobre contenidos de noticias, se encuentra un sistema de recomendación que se basa en el contenido de las mismas. Es un trabajo novedoso porque no se enfoca en el contenido general de las noticias para hacer el filtrado y la posterior recomendación, sino que primero se utiliza un "extractor de conceptos" en las páginas web de noticias para encontrar ideas más generales que puedan ser de ayuda para uno de los mayores problemas de los sistemas de recomendación: el inicio en frío. Una vez que hacen el procesamiento de la páginas web y las noticias contenidas en ella, estos 'conceptos' generados son comparados con los que se encuentren en el perfil del usuario construido por las noticias o sitios de noticias que él ha consultado anteriormente. [18]

Para ilustrar el método de filtrado colaborativo, es decir, aquél que busca comparar los intereses o actividad de los usuarios para encontrar patrones entre ellos, se tiene un sistema de recomendación que analiza las calificaciones y comentarios sobre restaurantes que dejan los usuarios en distintas páginas para determinar los intereses de los usuarios y las características de los restaurantes [19]. Otro trabajo que también hace uso del filtrado colaborativo es el que busca recomendar cursos electivos a estudiantes de una universidad específica. El sistema analiza el horario, preferencias (en cuanto a materias y profesores elegidos anteriormente) y los relaciona con el perfil de otros estudiantes, de esta manera puede recomendar cursos electivos que el estudiante podrá matricular. La ventaja de este enfoque es que al agrupar usuarios con características o intereses similares, brinda la oportunidad de que a un usuario determinado se le recomienden ítems que él mismo podría no haber considerado nunca, ofreciendo el beneficio extra de la novedad.[20]

En cuanto al enfoque híbrido se encontraron dos trabajos interesantes. El primero es tal vez una de las aplicaciones más frecuentes para los sistemas de recomendación: la música. Aunque en este trabajo no presentan un sólo método que combine ambos enfoques, sí muestran cómo son los resultados con cada uno de ellos. Así, las recomendaciones son realizadas basadas en unas características definidas y que son analizadas para agrupar la música, después según el perfil de los usuarios (construido con base en su comportamiento e intereses) se recomiendan los ítems más relevantes acorde a su historial. De igual manera muestran los resultados usando el filtrado colaborativo que agrupa música y también usuarios para recomendarles aquellos ítems que no se encontraban en su historial o intereses y que otros usuarios en su grupo sí los contenían.[21]

Otro sistema de recomendación que utiliza los dos enfoques (basado en contenido y filtrado colaborativo) busca recomendar medicamentos según el historial clínico de los usuarios. Esta es una aplicación importante y crítica porque afecta directamente el bienestar de las personas por lo que debe buscarse la mejor manera de implementarse y una metodología híbrida puede ser de gran utilidad. Su propuesta consiste básicamente en tres pasos: 1) análisis de las historias clínicas (donde se extraen aquellos términos o datos relevantes para el sistema como síntomas y medicamentos ya usados); 2) agrupamiento de los términos relacionados a síntomas o diagnósticos y aquellos relacionados a medicamentos; y 3) recomendación de los medicamentos apropiados. La precisión de los resultados obtenidos está alrededor del 80 % mostrando la utilidad de este enfoque híbrido y abriendo la posibilidad de mejoras al incorporar más datos relevantes sobre los usuarios.[22]

Aunque los métodos mencionados anteriormente son los más comúnmente utilizados para la implementación de sistemas de recomendación, no son los únicos. Otra variante importante y de gran relevancia en la actualidad es la minería de datos. Un buen ejemplo de esto se tiene en el sistema que usa los datos capturados por el GPS para recomendar los mejores puntos de recogida para un servicio de taxis, basándose en el origen y destino del usuario. La minería de datos se encarga de analizar grandes cantidades de datos para encontrar patrones o tendencias en diferentes contextos. En este trabajo, la información que se extrae de los datos suministrados por los GPS sirve para recomendar los puntos donde los usuarios pueden conseguir un servicio de taxi teniendo en cuenta factores importantes como las rutas, el tráfico, el tiempo del recorrido, el clima, entre otros[23]. Otra aplicación de este método se encuentra en la propuesta donde se recogieron los datos de los GPS de un número de vehículos que fueron alquilados por turistas en un período establecido y que con la minería de estos datos, pudo extraerse o encontrarse información que sería relevante para nuevos turistas que hicieran uso del alquiler de estos vehículos. Con estos datos se obtuvo información sobre sitios turísticos como restaurantes, museos, playas, etc. La minería de datos se está convirtiendo en una estrategia competitiva sumamente importante para las organizaciones con ánimo de lucro y que cuentan con una gran base de datos de sus operaciones y usuarios.[24]

Otras variantes o enfoques más especializados se ejemplifican en dos trabajos que hacen recomendaciones sobre maquillaje según el tipo de rostro[25] y hábitos alimenticios para tratar la malnutrición[26] Las recomendaciones de maquillaje usan el enfoque basado en reglas que especifica un conjunto de reglas que deben seguirse para clasificar y encontrar los patrones para realizar las recomendaciones. En éste primero se hace un análisis de los tipos de rostro con un número de imágenes usadas como conjunto de entrenamiento y la clasificación se realiza de acuerdo a diferentes características. Para determinar las reglas con las que se entrenaría el sistema, se entrevistaron expertos en maquillaje que dieron un conjunto de opciones básicas para cada tipo de rostro y maquillaje. Por su parte, el recomendador de hábitos alimenticios hace uso de una estrategia llamada “Honey Bees Mating Optimization” para hacer recomendaciones de alimentación a personas con riesgo de desnutrición. El sistema crea un perfil de usuario al monitorear los hábitos alimenticios de éste, después con ayuda de un árbol de toma de decisiones verifica si el usuario objetivo tiene riesgos de desnutrición. Las recomendaciones se hacen con el método mencionado y partiendo del perfil del usuario para analizar y sugerir los nutrientes que éste necesita para percibir mejoría en su salud.

2.2. Áreas de aplicación

2.2.1. Salud y estilo de vida

Tabla 2.1: Salud y vida

Artículo	**Descripción**	**Técnica implementada**	**Resultados destacados**
Rule-Based Facial Makeup Recommendation System [25]	El sistema de recomendación sugiere el mejor maquillaje dependiendo del tipo de rostro. Se usaron imágenes antes/-después para el entrenamiento y pruebas, y se estableció un conjunto de reglas con maquilladores expertos.	Basado en reglas	Los resultados se compararon con unos sitio web que procesa imágenes de rostros y agrega maquillaje a ellos. Se perciben buenos resultados con las recomendaciones del sistema.
Medication Recommendation System Based On Clinical Documents [22]	Sistema de recomendación para sugerir medicamentos para una enfermedad o síntoma específico. Se basa en el historial médico del paciente y de otros para realizar las recomendaciones.	Híbrido (Filtrado colaborativo y basado en contenido)	La precisión de los resultados experimentales está alrededor del 80
Lifestyle Recommendation System for Treating Malnutrition [26]	El sistema recomienda tipos de dieta a personas con riesgo de desnutrición cuya decisión se extrae del perfil de usuario construido por los hábitos alimenticios del usuario.	Honey Bees Mating Optimization	-

2.2.2. Educación

Tabla 2.2: Educación

Artículo	Descripción	Técnica implementada	Resultados destacados
Course Scheduler and Recommendation System for Students [20]	Plataforma online de programación y recomendación de cursos para los estudiantes de una universidad específica. Tiene en cuenta el horario, las materias y profesores elegidos anteriormente y las materias vistas por estudiantes en común para hacer las recomendaciones de los cursos electivos.	Filtrado colaborativo (FC)	Compararon los resultados obtenidos con diferentes algoritmos para hacer las recomendaciones y concluyeron que usando FC con una pequeña mejora que elimina cursos que se cruzan con sus horarios, se encontraron los mejores resultados.
Recommender Systems for University Elective Course Recommendation [27]	El sistema de recomendación busca recomendar cursos electivos que un estudiante podría tomar comparando su historial académico con el de otros estudiantes del mismo plan.	Filtrado colaborativo (usando Pearson y Alternating Least Square)	Basados en los resultados experimentales, encontraron que el algoritmo ALS presenta una precisión hasta de 86
Content-based Movie Recommendation within Learning Contexts [28]	Para relacionar un contexto educativo con recomendaciones de películas eligieron los documentales que aparecen en la base de datos del portal IMDb. Estas películas y los contenidos de estudio fueron clasificados teniendo en cuenta las categorías principales de Wikipedia. De esta manera, se realiza la correlación entre los dos objetos y la posterior recomendación.	Basado en contenido.	En los experimentos se analizó la relevancia y la relación entre los objetos recomendados. El 74

2.2.3. Turismo

Tabla 2.3: Turismo

Artículo	**Descripción**	**Técnica implementada**	**Resultados destacados**
A Restaurant Recommendation System by Analyzing Ratings and Aspects in Reviews [19]	El sistema crea perfiles para los usuarios y los restaurantes analizando las calificaciones y críticas que hacen los primeros en diferentes sitios web. Extraen la correlación entre estos tipos de feedback para encontrar los patrones y realizar las sugerencias.	Filtrado colaborativo	La precisión alcanza el 50
Demonstrator of a Tourist Recommendation System [24]	El sistema hace uso de los datos recolectados por GPS instalados en vehículos de alquiler para turistas, los procesa para encontrar y sugerir sitios turísticos que pueden ser de interés para otros turistas que hagan uso del servicio de alquiler.	Minería de datos	-

2.2.4. Transporte

Tabla 2.4: Transporte

Artículo	**Descripción**	**Técnica implementada**	**Resultados destacados**
TaxiHailer: A Situation-Specific Taxi Pick-Up Points Recommendation System [23]	El sistema de recomendación sugiere los mejores puntos de recogida para los usuarios dependiendo del origen, destino y hora de viaje. También tiene en cuenta factores como tráfico, clima, tiempo de espera, de viaje y ruta seleccionada.	Minería de datos, contexto	-
Context-Based Traffic Recommendation System [29]	El sistema se basa en la congestión de tráfico para recomendar las rutas en las que el usuario se tomará el menor tiempo posible, pero además pretende hacer que el viaje sea más placentero por lo que tiene en cuenta factores como restaurantes y sitios interesantes para sugerir mejores rutas.	Basado en contexto	-

2.2.5. Música

Tabla 2.5: Música

Artículo	Descripción	Técnica implementada	Resultados destacados
A Music Recommendation System Based on Music and User Grouping [21]	El sistema se basa en el perfil de usuario que se construye al revisar el historial del mismo y en las características extraídas de la música como duración, tono y volumen para clasificarlos y agruparlos. Las recomendaciones se realizan comparando usuario-música y también con los grupos de usuarios o canciones que se forman al hacer la clasificación.	Híbrido (Filtrado colaborativo y basado en contenido)	Al realizar los experimentos se encontró que el enfoque basado en contenido presenta mejores resultados que el filtrado colaborativo, una razón puede ser que el primero utiliza datos privados de los usuarios para mostrar las sugerencias por lo que estas se vuelven más personalizadas.
A Personalized Next-song Recommendation System using Community Detection and Markov Model [30]	Este sistema de recomendación tiene en cuenta las preferencias de los usuarios a corto y largo plazo. Para el largo plazo crea grupos de usuarios (como siempre se ha hecho) pero estos se dividen a su vez en 'comunidades' agregando un poco más de personalización. Para el corto plazo hacen uso de la técnica de Markov.	Modelo Markov	Al comparar los resultados con otros trabajos, encuentran que su propuesta se comporta muchísimo mejor respecto a la precisión a corto y largo plazo.
Design of Music Recommendation System Using Context Information [31]	El sistema clasifica los factores característicos de la música basado en ontología. El filtro de la música se hace con ayuda de un framework que trabaja basado en información de contexto.	Basado en contexto (Ontología)	Según los experimentos realizados y la retroalimentación recibida por parte de los usuarios en las pruebas, el 92

2.2.6. Películas y televisión

Tabla 2.6: Películas y televisión

Artículo	Descripción	Técnica implementada	Resultados destacados
A Smart Movie Recommendation System [32]	El sistema pide a los usuarios que ingresen sus géneros preferidos para encontrar la correlación con una lista de películas a las que se les da puntos dependiendo de su género. Además el sistema hace recomendaciones de películas que se encuentran en cartelera y se basa en la ubicación para mostrar información de los cines más cercanos al usuario.	Correlación de géneros. Basado en contenido	-
An Intelligent Recommendation System for TV Programme [33]	El sistema se conforma de 5 agentes: filtrado, recomendación, actualizador de perfil, reportes e interfaz. Los agentes de filtrado y recomendación son los que realizan las sugerencias basados en la similaridad entre las preferencias del usuario y las características de los programas.	Basado en contexto.	Los experimentos realizados encontraron que el sistema puede aprender de las preferencias del usuario aunque éstas cambien abruptamente, pero para esto es necesario una retroalimentación explícita por parte del usuario.

Artículo	**Descripción**	**Técnica implementada**	**Resultados destacados**
The MovieOracle Content Based Movie Recommendation [34]	Para clasificar las películas por palabras o conceptos, se recogen los diálogos completos de las mismas (en forma de sus subtítulos en un archivo de texto) y de esta manera se extrae su categorización. Así se pueden encontrar relaciones entre películas por sus diálogos o se pueden ingresar palabras claves y generar recomendaciones. También fue posible generar las recomendaciones al subir el contenido de un libro al framework construído.	Basado en contenido.	Los experimentos realizados arrojaron recomendaciones de películas similares (como sagas) y sin necesidad de la interacción de un ser humano.
Content-Based Filtering Recommendation Algorithm Using HMM [35]	Los datos son descargados desde MovieLens, estos contienen la información más relevantes sobre una película pero el atributo que realmente se usó fue el o los géneros a los que pertenece la misma. Usando K-means se generan los clusters relacionados a los 19 géneros contenidos en la base de datos. Usando el algoritmo HMM se efectua un número determinado de iteraciones para re-organizar los clusters y obtener una clasificación más precisa. Las películas que se recomiendan pertenecen al mismo cluster (o género).	Basado en contenido (Hidden Markov Model)	Con los experimentos y pruebas realizados se obtuvo una precisión significativamente más alta al ser comparada con un algoritmo basado en el modelo espacio vectorial (VSM).

2.2.7. Noticias

Tabla 2.7: Noticias

Artículo	**Descripción**	**Técnica implementada**	**Resultados destacados**
CONCERT: A Concept-Centric Web News Recommendation System [18]	El sistema de recomendación se basa en el contenido de las noticias en diferentes páginas web, extrayendo las ideas principales de éstas como 'conceptos' y formando un perfil de usuario con los sitios que él ha visitado anteriormente. Las recomendaciones se realizan en tiempo real comparando los 'conceptos' de las noticias con los del perfil de usuario.	Basado en contenido	-
A Semantic Content Based Recommendation System for Cross-Lingual News [36]	El sistema procesa las noticias en Bengalí pero hará recomendaciones de noticias en inglés con ayuda de la web semántica y formando una ontología automáticamente.	Basado en contenido (ontología)	El sistema de recomendación se comporta de mejor manera que uno basado en palabras claves por la adición de la semántica.

2.3. Bibliotecas

Tabla 2.8: Bibliotecas

Artículo	**Descripción**	**Técnica implementada**	**Resultados destacados**
A Personalized Time-Sequence-Based Book Recommendation Algorithm for Digital Libraries [37]	El algoritmo recomienda libros a los universitarios teniendo en cuenta sus carreras profesionales para que las sugerencias sean más precisas. El sistema trabaja con la información de préstamos y de circulación de los libros.	Filtrado colaborativo	Los resultados que obtuvieron, al ser comparados con los de otros trabajos, no son tan relevantes. Una razón importante que influye en ellos es que no poseen una buena cantidad de datos para llevar a cabo los experimentos.
Book Recommendation Based on Book-Loan Logs [12]	Primero analizan el historial de préstamos del usuario para clasificarlo o perfilarlo. Las recomendaciones se hacen de dos maneras: por el método tradicional de filtrado colaborativo y otro basado en probabilidad teniendo en cuenta la carrera profesional del usuario.	Filtrado colaborativo y probabilidad	Aunque el filtrado colaborativo recomienda más novedades, se encuentra que el método probabilístico propuesto en el artículo realiza mejores recomendaciones acordes a la carrera profesional del usuario.
Book Recommendation Using Machine Learning Methods Based on Library Loan Records and Bibliographic Information [7]	Usaron diferentes métodos para las recomendaciones como SVM y dos paquetes de R (Adaboost y Random Forest), así como diferentes combinaciones de características de los libros, por ejemplo: títulos, préstamos, año de publicación y frecuencia de préstamo.	Support Vector Machine, Adaboost, Random Forest	Los experimentos mostraron que SVM fue el método que presentó mejores resultados, incluso al ser comparado con el algoritmo de Amazon.

Artículo	**Descripción**	**Técnica implementada**	**Resultados destacados**
Online Book Recommendation System [38]	Los usuarios deben registrarse en el sistema y en ese momento deben proveer la retroalimentación sobre sus gustos y preferencias en libros o categorías. Con esta información el sistema realiza las recomendaciones complementandolo con el perfil construido por otros usuarios.	Filtrado colaborativo	Los experimentos se realizaron con retroalimentación de usuarios de prueba. La calidad de las recomendaciones alcanzó el 77
Collaborative Book Recommendation Based on Readers' Borrowing Records [8]	Para realizar las recomendaciones hacen una mezcla de métodos usados en el enfoque del filtrado colaborativo. La información que utilizan para el procesamiento son los préstamos de los libros, y al no tener una calificación por parte de los usuarios, transforman el tiempo de préstamos de los libros en un promedio de calificación.	Filtrado colaborativo	En términos de precisión, una de las mezclas de técnicas que implementaron se comporta de mejor manera que las técnicas individuales. Con una de las últimas, podría predecirse el tiempo de préstamo de los libros por parte de un usuario.
A Novel Approach for Book Recommendation Systems [39]	Para recopilar información sobre el usuario, el sistema analiza los comentarios y calificaciones de los libros que ha comprado anteriormente. Así se incluyen los libros y su puntaje correspondiente en el perfil de usuario. Para recomendar otros libros del interés del usuario, proponen un algoritmos que encuentra las asociaciones entre libros de una manera más rápida que otros algoritmos usados comúnmente.	Minería de datos.	Al comparar el algoritmo tradicional y el que se propone en el artículo, encuentran que con un número reducido de transacciones ambos se comportan de la misma manera. Pero a medida que las transacciones aumentan, el algoritmo propuesto presenta un mejor comportamiento al tomar menos tiempo para realizar escaneos de toda la información y mostrar las recomendaciones.

Artículo	**Descripción**	**Técnica implementada**	**Resultados destacados**
Web-based Personalized Hybrid Book Recommendation System [40]	El sistema contiene la información de los usuarios, los libros y la calificación de estos. La primera fase del filtrado es un híbrido entre el enfoque basado en contenido y el colaborativo, después se pasa a un filtrado demográfico y se incorpora también información obtenida con ayuda de la minería web. Finalmente se obtienen las recomendaciones personalizadas para cada usuario.	Híbrido (colaborativo, contenido, demográfico,minería web)	-
Use of Library Loan Records for Book Recommendation [6]	Se usaron tres enfoques para determinar qué tipo de método es más eficiente a la hora de recomendar libros basándose en la información de los préstamos. Los datos usados para encontrar los patrones y para el filtrado fueron el ID del usuario, ID del material que ha prestado anteriormente, información bibliográfico sobre los libros y la fecha de devolución.	Filtrado colaborativo, reglas de asociación, algoritmo de Amazon	Los experimentos realizados encontraron que el método más ineficiente es el colaborativo (que además pone en riesgo la privacidad de los usuarios si es mal implementado), seguido por las reglas de asociación y concluyendo que el algoritmo de Amazon presenta los mejores resultados en la precisión de las recomendaciones.
A Data Mining Approach to New Library Book Recommendations [41]	Utilizando información demográfica de los usuarios y atributos de los libros, se categorizan ambos en diferentes tipos para encontrar asociaciones basados en reglas de asociación generalizadas. Busca mejorar el problema al recomendar libros nuevos que no cuentan con calificación o préstamos.	Minería de datos (reglas de asociación de perfil generalizadas)	-

Artículo	Descripción	Técnica implementada	Resultados destacados
A Novel FAHP Based Book Recommendation Method by Fusing Apriori Rule Mining [42]	El algoritmo apriori es implementado para encontrar las asociaciones entre los libros en circulación. Para tener en cuenta características o atributos de los usuarios y de los libros, hacen uso del framework FAHP que define una estructura de tres capas para clasificar los libros.	Híbrido (Fuzzy Analytical Hierarchy Process (FAHP) y Minería de reglas apriori)	Con una medida de precisión implementada por ellos y comparando los resultados con un algoritmo de filtrado colaborativo, encuentran que la propuesta presenta mejores resultados a nivel de precisión en las recomendaciones y además realiza un número mayor de las mismas.
An enterprise-friendly book recommendation system for very sparse data [43]	Como primera fase se usa k-means para formar clusters de usuarios y libros. El siguiente paso fue formar una "matriz bicluster" con los ratings de los usuarios a libros. Para los biclusters se usó la similaridad entre los valores de las celdas de la matriz. Las recomendaciones se hicieron teniendo en cuenta el "vecino más cercano".	Minería de datos - Algoritmo biclustering (híbrido)	Usando valores de error MAE y aplicando el algoritmo en diferentes tipos de datasets, se encontraron valores de error relativamente bajos cuando el dataset contaba con mucha dispersión. Los resultados en el resto de datasets también fueron relativamente buenos.
Book Recommendation System Based on Collaborative Filtering and Association Rule Mining for College Students [44]	El sistema busca recomendar libros de texto académicos a usuarios (potenciales compradores). Usa el filtrado colaborativo para seleccionar los libros más pertinentes a cada usuario, y la minería por reglas de asociación es un paso más para refinar el filtrado basándose en los rangos de precios de los libros y el autor.	Híbrido (filtrado colaborativo y minería por reglas de asociación)	-

Artículo	**Descripción**	**Técnica implementada**	**Resultados destacados**
Book Recommendation System for Digital Library Based on User Profiles by Using Association Rule [45]	Haciendo un análisis de los préstamos de los usuarios para formar su perfil, y de las categorías de los libros se crea un modelo basado en reglas de asociación para realizar las recomendaciones en busca de facilitar las búsquedas de los usuarios y mejorar los resultados mostrados.	Minería de datos - reglas de asociación	Se realizaron experimentaciones con diferentes reglas para medir la confianza de cada una, aquellas que sobrepasan un mínimo determinado se aceptan. En cuanto a la matriz de evaluación, se encontraron porcentajes de precisión hasta del 90
Book Recommendation System through Content Based and Collaborative Filtering Method [11]	Extraen la información relevante de la base de datos(minería), forman clusters de los libros en géneros y subgéneros (clustering), así le aplican el filtrado basado en contenido de acuerdo a las preferencias del usuario, a estos resultados aplican filtrado colaborativo basado en la calificación y opiniones de los usuarios y finalmente se muestran los resultados de manera descendente dependiendo de la clasificación.	Híbrido (filtrado colaborativo, basado en contenido y minería por reglas de asociación)	-
Collaborative Book Recommendation System using Trust based Social Network and Association Rule Mining [46]	Al crear el perfil del usuario con sus calificaciones y transacciones anteriores sobre libros, pasa a procesarse su lista de amigos en x red social para encontrar aquellos con más confianza y aplicar el filtrado colaborativo. Con las reglas de asociación se intenta que la calidad de las recomendaciones se incremente.	Híbrido (filtrado colaborativo y minería reglas de asociación)	Al realizar los experimentos y comparar los resultados con la técnica de filtrado colaborativo, se encuentra que la propuesta presenta mejores niveles de precisión sobre las recomendaciones.

Artículo	Descripción	Técnica implementada	Resultados destacados
Library Book Recommendations Based on Latent Topic Aggregation [47]	Al crear el perfil de usuario con sus transacciones históricas, se usa el número de catálogo de los libros en vez de su id o título, y al aplicarle el método LDA se espera encontrar una recomendación personalizada de interés al usuario.	Filtrado colaborativo (latent dirichlet allocation)	Los experimentos mostraron 60 % de satisfacción del usuario, un buen resultado que podría mejorarse, por ejemplo incluyendo no sólo los top n libros generados por la recomendación sino agregar además algunos del final de la lista.
Library Management System Based on Recommendation System [13]	Se plantea un diseño de la estructura, módulos funcionales e interfaz con la que debe contar un sistema de recomendación de libros con la técnica de filtrado colaborativo. La arquitectura consta de 3 capas: capa de datos, capa de procesamiento y capa de vista.	Filtrado colaborativo	-
Online Book Recommendation System by using Collaborative filtering and Association Mining [10]	Con la propuesta híbrida buscan solucionar el problema de dispersión de datos. Primero se usa el filtrado colaborativo para encontrar similitudes entre libros. Las reglas de asociación se usan para llenar las calificaciones que falten y tratar el problema de dispersión. Para las recomendaciones se usa igual filtrado colaborativo entre usuarios.	Híbrida (filtrado colaborativo y minería reglas de asociación)	Para las pruebas usaron el valor absoluto medio y se encontraron mejores resultados a medida que realizaban un mayor número de recomendaciones.

Artículo	**Descripción**	**Técnica implementada**	**Resultados destacados**
Content Based Book Recommending Using Learning for Text Categorization [48]	Primero extraen la "base de datos" desde la información disponible en la página de Amazon con la información de los libros y además los comentarios que los usuarios han proveído para estos. Esta información se procesa para encontrar las palabras o conceptos más representativos de cada libro. Con la información obtenida se seleccionan usuarios para que califiquen un número determinado de libros y categorizar el perfil de usuario para poder efectuar las recomendaciones.	Basado en contenido	Con los experimentos iniciales obtienen buenos resultados al efectuar recomendaciones "Top-3" con un porcentaje cerca del 90 % de precisión. Se considera un muy buen resultado puesto que no se está utilizando información de ningún otro usuario, es decir, ningún método de filtrado colaborativo que se dice es más eficiente.
Content-Based Recommendation Services for Personalized Digital Libraries [49]	El framework propuesto construye un perfil del usuario semántico que contiene sus intereses representados en conceptos claves obtenidos de una serie de documentos. Para la recomendación se realiza el paralelo entre los conceptos incluídos en el perfil del usuario y aquellos que describen los ítems o libros.	Basado en contenido	-

Capítulo 3

Alcance de la propuesta

3.1. Alcance de la propuesta

El prototipo de sistema de recomendación web se implementó con base en la información que fue suministrada por la Biblioteca Mario Carvajal y una base de datos externa en inglés. Al obtenerse los datos del catálogo de la biblioteca, se aplicaron técnicas de recomendación que se basan en el contenido de aquella información. Los datos de préstamos de los usuarios de la biblioteca se utilizaron para mostrar un perfil de usuario donde ellos pueden obtener resultados más personalizados.

Por el tiempo con el que se cuenta para el desarrollo del proyecto, se seleccionaron solamente dos técnicas o algoritmos de recomendación para ser modelados e implementados en el problema propuesto.

La implementación del prototipo del sistema web permite dos tipos de búsqueda: si se realiza la consulta de un libro (sobre los datos de la base de datos externa), la aplicación despliega un listado de coincidencias donde el usuario puede seleccionar un elemento para que el prototipo le arroje una lista de libros similares recomendados. Con las transacciones de los usuarios de la biblioteca Mario Carvajal, un usuario puede ingresar su código y obtener una lista de recomendaciones personalizadas basadas en alguno de sus préstamos anteriores. Al seleccionar alguno de los libros de la lista de recomendaciones, se muestra una vista previa de los datos más relevantes del elemento.

Todas las operaciones se hacen en un prototipo de sistema web externo a la plataforma de la biblioteca y utilizando la información suministrada por el personal de la biblioteca además de una base de datos adicional sobre libros en inglés, no se harán ejecuciones en tiempo real. El prototipo queda a disposición de la biblioteca para su posterior integración con la plataforma oficial si así lo desean.

Capítulo 4

Modelo

4.1. Técnicas basadas en contenido

4.1.1. TF-IDF

La técnica de procesamiento de lenguaje natural $tf - idf$ es uno de los métodos más conocidos que se utilizan en el procesamiento o minería de textos y evalúa la importancia de las palabras o frases de un documento.

El término TF ($termfrequency$) hace referencia a la frecuencia de una palabra en un documento o texto y se define como $tf(t,d) = \sum_{x \in d} fr(x,t)$. Donde $fr(x,t)$ es una función definida como:

$$fr(x,t) = \begin{cases} 1, \text{ if } x = t \\ 0, \text{ otherwise} \end{cases}$$

Donde $fr(t,d)$ retorna como resultado el número de veces que el término t aparece en el documento d.

Uno de los problemas que presenta TF es que beneficia mucho los términos frecuentes y "castiga" a los términos raros (que en la mayoría de casos son los más significativos para el documento). La premisa básica es que un término que es frecuente en muchos documentos no necesariamente es un buen discriminador (tiene sentido en experimentos empíricos del estado del arte).

La estrategia tf-idf soluciona el problema anterior al restarle importancia a los términos de más frecuencia y aumentando la de los términos significativos (un término que aparece 10 veces más que otro no es 10 veces más importante). Tf-idf calcula la importancia de un término u oración para un documento dentro de una colección de éstos definiendo los parámetros locales como los valores del término dentro de su propio documento y los parámetros globales como los valores del término dentro de toda la colección de documentos.

El resultado de obtener el valor tf-idf de un documento es un vector que posteriormente se normaliza para evitar que los documentos más extensos (aquellos con más palabras y/o frecuencia de palabras) se consideren de más importancia. Esta normalización se realiza con la siguiente fórmula:

$$v = \frac{\vec{v}}{||\vec{v}||_p}$$

Donde $\vec{v}$ es el vector y $||\vec{v}||_p$ es la norma euclidiana con $p = 2$ [50] y que da como resultado el vector normalizado con valores entre 0 y 1.

El término IDF ($frecuenciadeterminoinverso$) busca la importancia de un término de manera global, inspirándose en la propiedad del logaritmo [51]:

$$idf(t) = log\frac{|D|}{1 + |\{d : t \in d\}|}$$

Siendo $|D|$ el número total de documentos y $1 + |\{d : t \in d\}|$ el número de documentos donde aparece el término t. En este término se suma un 1 para evitar una división por cero.

En sintesis, la fórmula tf-idf(t)=tf(t, d)xidf(t) calcula la importancia del término local y globalmente.

Para representar los valores $tf - idf$ de todos los documentos, los vectores se almacenan en una matriz $_{\mathrm{M}}$tf-idf donde cada fila representa los documentos de la colección y las columnas son cada uno de los términos contenidos en los documentos.

Finalmente se realiza la normalización de la matriz usando la norma 2 [50] así:

$$_{\mathrm{M}}tfidf = \frac{_{\mathrm{M}}tfidf}{||_{\mathrm{M}}tfidf||_2}$$

4.1.2. Similaridad coseno

Cuando se tiene una matriz de vectores a la que se le quiere encontrar la similaridad entre ellos, el método de similaridad coseno entra en juego al calcular el coseno del ángulo entre dos vectores. Cuando los vectores han sido normalizados, la métrica de la distancia toma como referencia el ángulo entre vectores y es una medida de la orientación y no de la magnitud de ellos.

Para encontrar el coseno del ángulo se toma el cálculo del producto punto de los vectores y se despeja el coseno del ángulo así:

$$\vec{a}.\vec{b} = ||\vec{a}||||\vec{b}||cos\theta \rightarrow cos\theta = \frac{\vec{a}.\vec{b}}{||\vec{a}||||\vec{b}||}$$

Con los vectores normalizados, la magnitud de cada uno es igual a 1, por lo que el cálculo del coseno del ángulo se reduce a simplemente hallar el producto punto de los vectores.

Con un ángulo pequeño se tiene un coseno de mayor valor, denotando más similaridad entre los vectores(items).

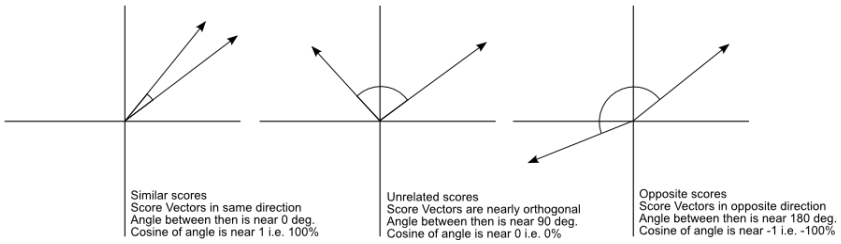

Figura 4.1: Ejemplo de similaridad coseno

4.2. Técnicas basadas en filtrado colaborativo

4.2.1. SVD

La Descomposición en valores singulares es una técnica de factorización de matrices donde se busca reducir la dimensión de una matriz hallando tres matrices que al ser multiplicadas entre sí dan como resultado la matriz inicial.

$$\underset{m \times n}{\hat{X}} \approx \underset{m \times r}{U} \; \underset{r \times r}{S} \; \underset{r \times n}{V^{\mathsf{T}}}$$

Figura 4.2: Descomposición de matrices en valores singulares

Si se tiene una matriz X que relaciona items-usuarios y los valores de la matriz son las calificaciones de usuarios sobre estos items, ésta se puede descomponer en las siguientes tres matrices:

- La matriz U que es una matriz singular izquierda que relaciona a los usuarios y los factores latentes.
- La matriz S es una matriz diagonal que describe la fuerza/importancia de cada factor latente.
- La transpuesta de la matriz V es una matriz singular derecha que relaciona la similaridad entre items y los factores latentes.

Al hablar de factor latente se hace referencia a algún concepto o característica general que pueden tener los usuarios o items.

Para lograr entonces una aproximación de la matriz inicial pero con una dimensión reducida, se observan los factores latentes de mayor valor en las matrices obtenidas de la descomposición y se eligen las columnas correspondientes en las matrices item-factorlatente y user-factorlatente.

4.2.2. Coeficiente de Correlación de Pearson

Es una medida de la relación lineal entre dos variables cuantitativas.

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Donde σ_{XY} es la covarianza entre X y Y, y σ_X σ_Y son las desviaciones estándar de X y Y respectivamente. La asignación en una banda se representa en una matriz de n operadores por c canales, donde la posición i, j representa la asignación del operador i en el canal j.

Dependiendo del resultado obtenido por el coeficiente de correlación se puede deducir la relación entre las variables así:

- Si el coeficiente es igual a 1, quiere decir que las variables tienen una correlación positiva perfecta, cuando una aumenta la otra lo hace en la misma proporción.
- Si el coeficiente es igual a -1, quiere decir que las variables tienen una correlación negativa perfecta, cuando una aumenta la otra disminuye en la misma proporción.

- Entre más cerca esté el valor del coeficiente a 1, la correlación entre las variables es mayor. Si el valor es más cercano a 0, las variables no poseen una correlación lineal tan fuerte.

Cuando se tiene una matriz que busca la correlación del coeficiente de Pearson entre sus elementos, la diagonal siempre será 1 (o muy cercano a 1 si existen problemas de redondeo) pues estos campos indica la correlación de un elemento con él mismo.

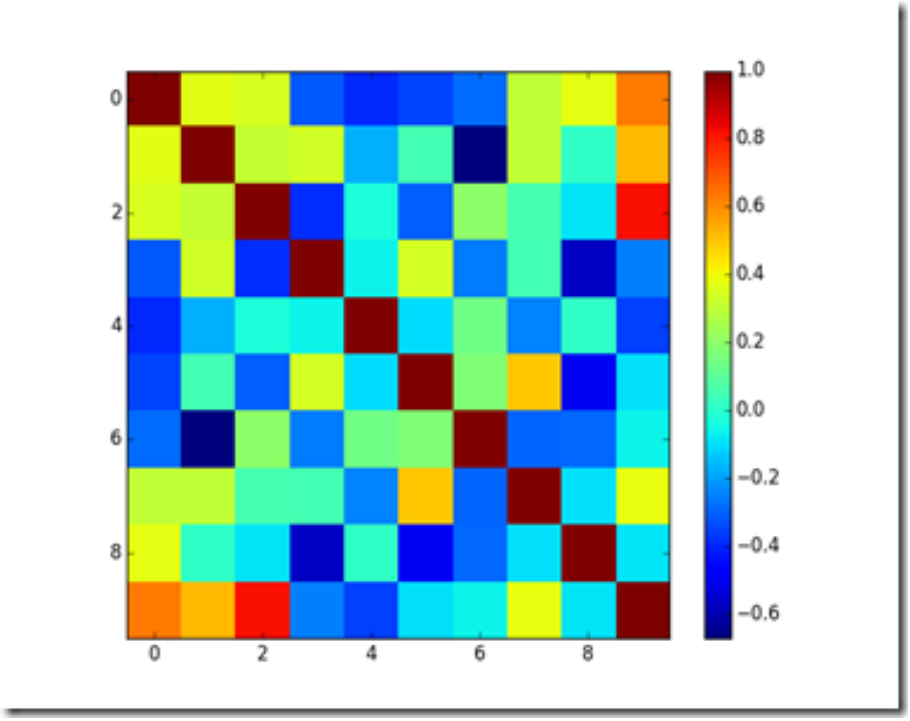

Figura 4.3: Ejemplo de matriz de correlacion

4.2.3. Vecinos más cercanos

La técnica de los k vecinos más cercanos es una de las más simples, populares y relativamente eficientes en términos de clasificación. Dentro de las características presentes en este método se tiene que es no-paramétrico, lo que quiere decir que no tiene en cuenta ningún dato sobre la distribución de la información ni tampoco realiza suposiciones sobre ella. Otra de sus características es que la fase de entrenamiento no es explícita porque lo que no se presentan generalizaciones. En esta fase se determinan los k grupos que se identifican por un atributo o característica de los elementos pertenecientes a cada uno de ellos.

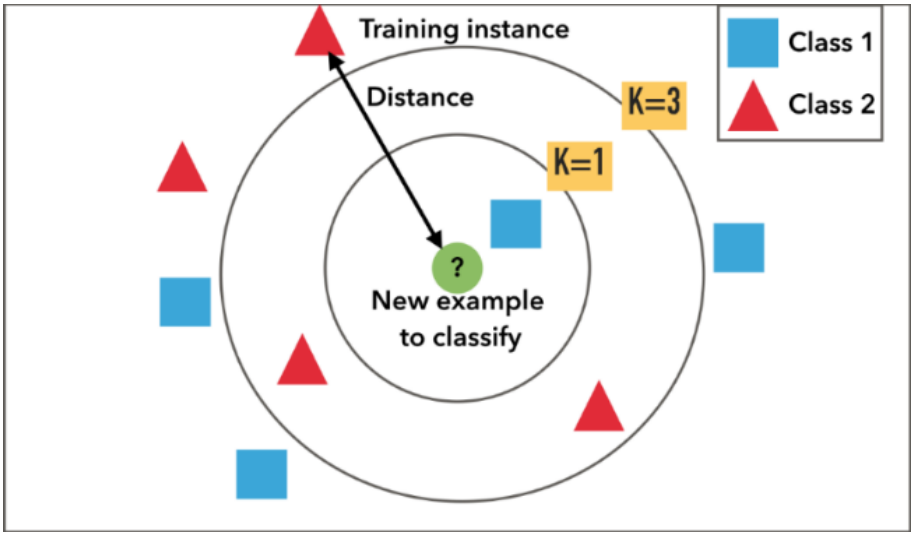

Figura 4.4: Clasificación de un nuevo elemento con vecinos más cercanos

La premisa básica bajo la idea de knn se basa en la característica de similaridad(distancia), cuando se tiene un nuevo elemento (perteneciente a los datos de prueba) su clasificación se hace dependiendo de sus vecinos. Son los vecinos los que "votan" para que el nuevo elemento sea incluido dentro de su clase, por lo que al estar cerca de más vecinos de una misma clase, se da lugar a la introducción del elemento dentro de su grupo.

La medida de distancia para calcular la similaridad utilizada en esta técnica es la similaridad del coseno que fue explicada anteriormente.

Capítulo 5

Metodología

5.1. Descripción de SCRUM

La metodología utilizada para el desarrollo del proyecto fue SCRUM. Esta técnica hace uso de unos roles donde se definen las tareas y responsabilidades que cada uno de los involucrados en el proyecto deben cumplir y llevar a cabo, así como actividades diarias y semanales donde se revisa el avance del proyecto.

Para este trabajo se determinaron pocos roles considerando que no se cuentan con muchas personas implicadas en la elaboración del proyecto, estos se pueden enunciar así:

- Como *Dueño del producto (Product owner)* se tiene al director de trabajo de grado quien cumplirá las funciones de .[en]laceçon el cliente/usuario, ayudando a definir claramente los requerimientos e historias de usuario y priorizando aquellas que deben entregarse en cada ciclo, pues es quien más entiende lo que debe representar el producto final.
- El *SCRUM Master* en este caso también es el director del trabajo de grado, quien se encargará específicamente de liderar y guiar al equipo de desarrollo en el cumplimiento de los requerimientos del proyecto y de los lineamientos de SCRUM.
- El *equipo de trabajo* se conforma solamente por una estudiante a quien pertenece el trabajo de grado y que lleva a cabo las funciones de desarrollo y terminación del producto.

Dentro de las actividades propias de SCRUM se tienen:

- Definición de requerimientos e historias de usuario al inicio del proyecto.
- Por el tipo de proyecto, se realizaron reuniones semanales donde se evaluaba el avance hasta ese momento y las dificultades que se presentaron. Además se priorizaban las tareas a seguir en la siguiente semana. Estas actividades hacen referencia al *scrum diario* y al *sprint*.

Capítulo 6

Desarrollo

6.1. Desarrollo del proyecto

La primera fase del proyecto fue el preprocesamiento de la información de la base de datos que proporcionó la biblioteca Mario Carvajal y de la base de datos externa utilizada para la recomendación del filtrado colaborativo que puede encontrarse en [52].

El desarrollo del proyecto se realizó utilizando el framework para aplicaciones web Django en su versión más reciente, junto con Python 3.6 y las librerías sklearn, numpy, scipy, pandas y psycopg2. Esta selección se hizo teniendo en cuenta el gran número de aplicaciones referentes a la ciencia de datos, machine learning y minería de datos con las que cuentan Python gracias a la integración que permite con las librerías mencionadas.

La estructura implementada en el proyecto de Django se puede definir de la siguiente manera:

- Una app 'accounts' para los usuarios con transacciones en la biblioteca Mario Carvajal, que aunque no se tiene más información específica sobre ellos, es necesaria para la implementación del inicio de sesión en el prototipo.
- Una app 'biblioteca' que maneja la información suministrada en la base de datos de la biblioteca Mario Carvajal y la implementación de las técnicas de recomendación basadas en contenido.
- Una app 'bxlibrary' que maneja la información de una base de datos externa[52] con información sobre libros, usuarios y ratings, además de la implementación de las técnicas de filtrado colaborativo

6.1.1. App Accounts

Dentro de la información obtenida de la base de datos de la biblioteca Mario Carvajal, el único registro que hace referencia o suministra información sobre los usuarios es el código de usuario en la tabla "Transacciones". Tomando como base el sistema OPAC con el que cuenta la Universidad del Valle, en el cual el inicio de sesión se realiza con el código de usuario, se implementa de igual manera un módulo de logueo para los usuarios en el que el username y contraseña es su código de usuario. El prototipo tampoco cuenta con una opción de registro para nuevos usuarios.

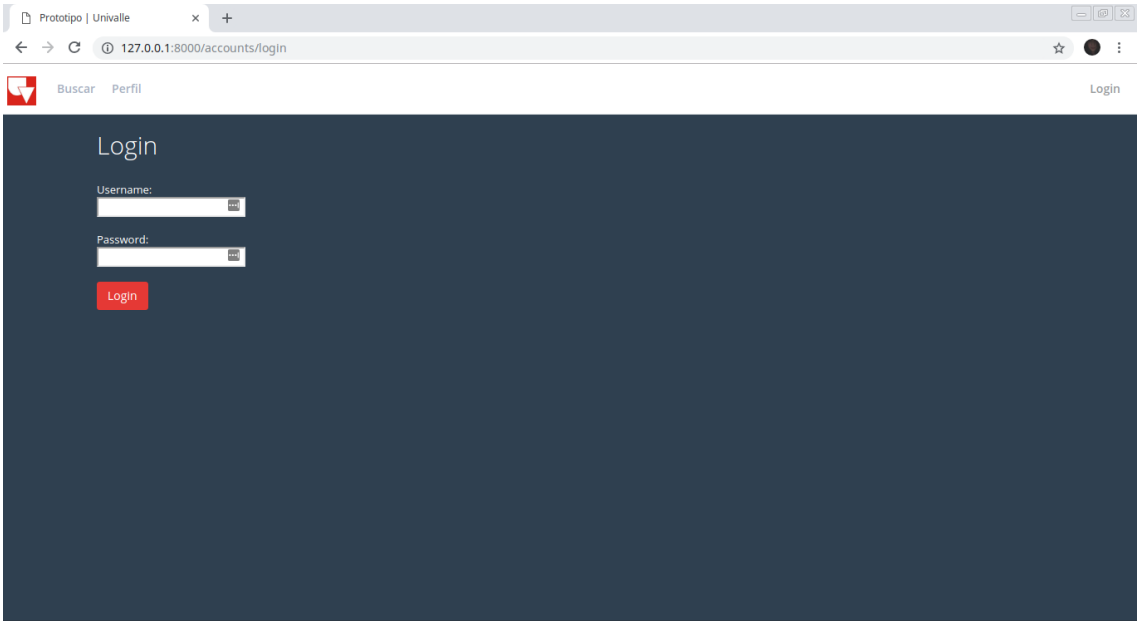

Figura 6.1: Pantalla de logueo

Aunque se tiene claro que esto representa un problema para la privacidad y seguridad de los usuarios, se realiza como una medida provisional teniendo en cuenta que se está desarrollando un prototipo que quedará en manos de la biblioteca Mario Carvajal.

6.1.2. App Biblioteca

Inicialmente, la base de datos contaba aproximadamente con 370.000 registros de libros, 330.000 autores y alrededor de 14.000.000 de registros de transacciones, entre otras tablas que conforman la información general del material bibliográfico.

Después de definir los requerimientos y un bosquejo del prototipo (se pueden observar en los anexos), se planteó la recomendación basada en contenido para este módulo. Cuando un usuario inicia sesión, se le redirige a su "perfil" de usuario que muestra las transacciones que éste ha realizado sobre/con el material bibliográfico. El usuario puede seleccionar cualquiera de los elementos listados para obtener recomendaciones de items similares a éste.

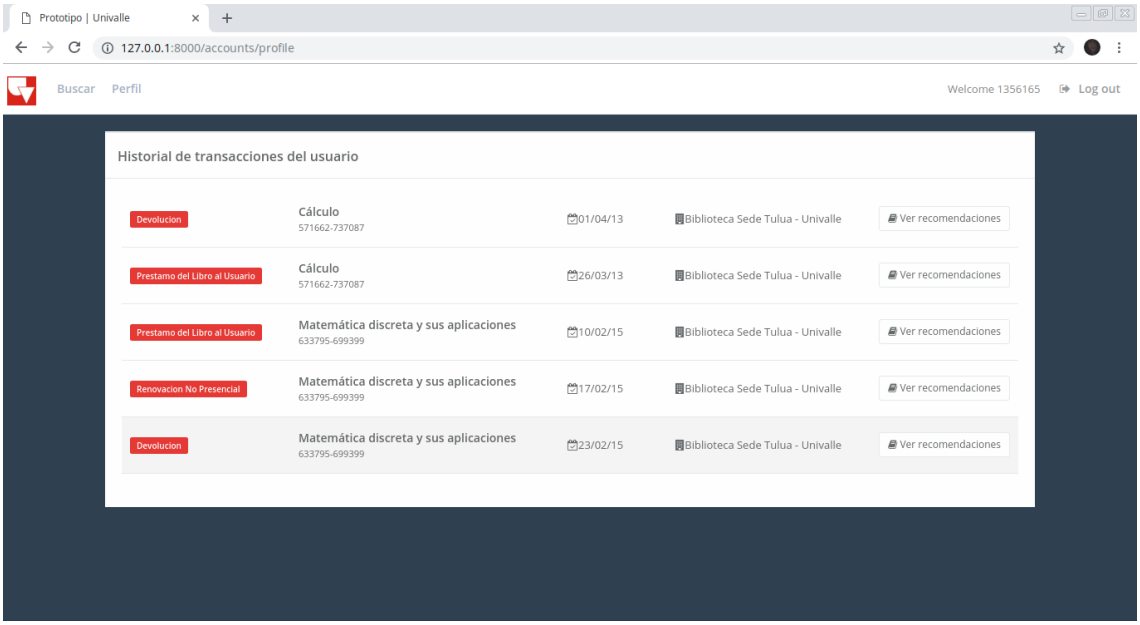

Figura 6.2: Perfil de usuario

Al realizar pruebas previas de las técnicas basadas en contenido sobre el total de información de la base de datos, se concluye que no es posible trabajar con un volumen tan extenso de información, un equipo de cómputo tan limitado y con algoritmos "simples" como los implementados. Por esta razón se eligen cierto número de registros al azar para formar una nueva base de datos de menor tamaño que permita la ejecución de los métodos de recomendación de contenido.

Así, la nueva base de datos cuenta con aproximadamente 27.000 títulos, 330.000 autores y 950.000 transacciones. Estas cantidades se obtuvieron después de realizar diferentes pruebas y son el número límite con el que puede trabajar relativamente bien el equipo de cómputo disponible.

6.1.3. Técnicas de recomendación basadas en contenido: similaridad del coseno, linear kernel

La función implementada para la recomendación basada en contenido consta de los siguientes pasos:

1. El primer paso es realizar una consulta a la base de datos que realice múltiples "join" con las tablas de títulos(libros), autores, notas y materias para obtener la información y características que describe a los libros.

2. El resultado de la consulta anterior pasa a convertirse en un DataFrame que es uno de los tipos de estructura de datos que maneja la librería de pandas [53]

3. Cuando se tiene la matriz (dataframe) de libros con toda la información correspondiente, se define la columna "descripción" que contiene las materias (categorías como género, temática, etc) y notas del libro (generalmente es el contenido, resumen o información adicional del material).

4. En este punto se hace uso de la técnica tf-idf llamando una función del paquete sklearn [54], que formará una matriz de vectores con los términos y pesos de la descripción de los libros contenidos en la matriz inicial.

5. Con la matriz tf-idf se busca entonces el nivel de similaridad que existe entre todos los libros, haciendo uso de la función del paquete sklearn 'cosine_similarity' que calcula el coseno del ángulo entre los vectores tf-idf de los libros que se estén comparando. En el capítulo del modelo del proyecto se realiza una explicación más detallada sobre este método.

6. Con la misma matriz tf-idf se usa también otra función del paquete sklearn llamada 'linear_kernel' que para este caso (donde se tiene una matriz tf-idf con vectores normalizados) es equivalente a la función 'cosine_similarity' pero produce resultados de forma más rápida al realizar simplemente el producto punto entre los vectores de los libros que se comparen.

7. Las funciones de los dos anteriores retornan cada una la matriz con los valores de similaridad entre todos los libros.

8. Con ayuda de los métodos que manipulan los dataframe, se localiza la posición del libro objetivo (aquél al que se le quieren encontrar libros similares) y con este índice pueden obtenerse los mayores valores de similaridad con otros libros.

9. Finalmente se retornan las listas con las puntuaciones, identificadores e información general sobre los libros recomendados.

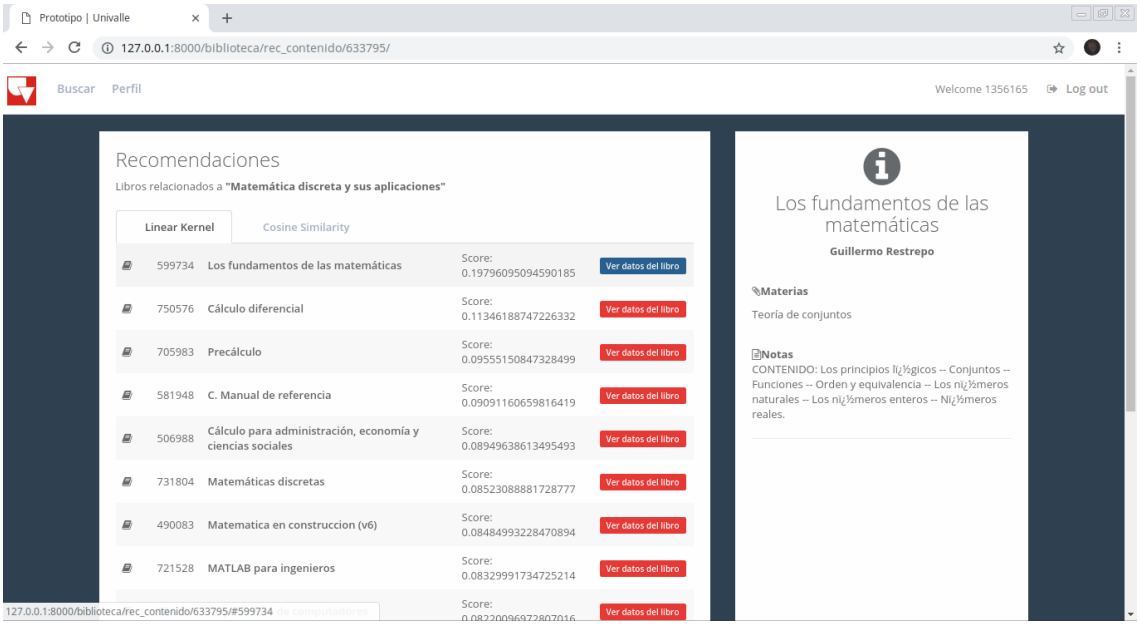

Figura 6.3: Recomendaciones basadas en contenido

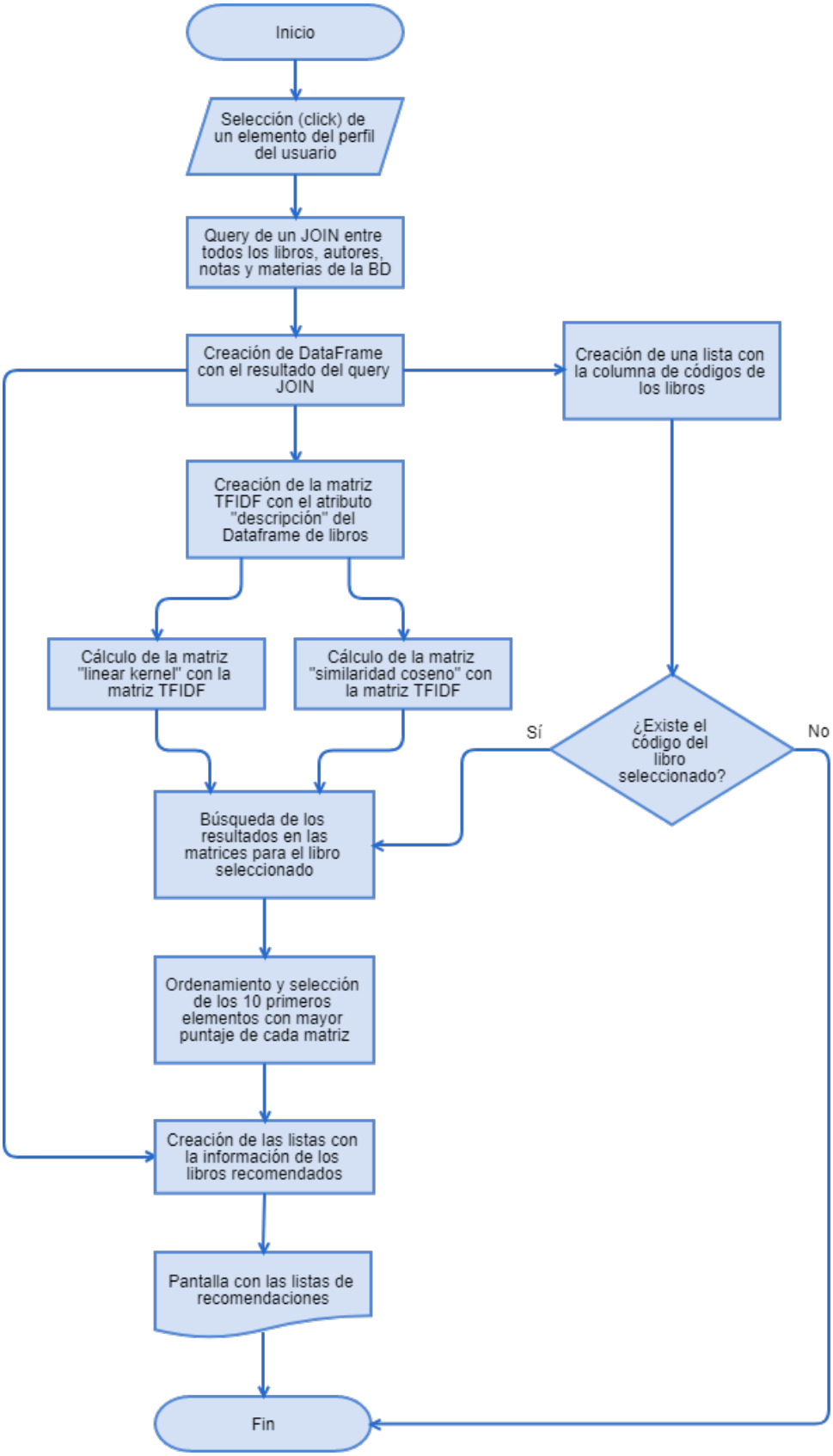

Figura 6.4: Diagrama de flujo para las recomendaciones basadas en contenido

6.1.4. App BxLibrary

Cuando se obtuvo la información de la base de datos de la biblioteca Mario Carvajal, se pudo observar que los registros no tenían ningún tipo de calificación o atributo que denotara una medida de interacción de los usuarios con el material bibliográfico. Por esta razón se decidió utilizar una base de datos externa de libros con ratings.

Según el alcance definido para el proyecto, se debía implementar una de las técnicas de recomendación para las consultas que los usuarios realizaban en el prototipo, de esta manera se determinó en las historias de usuario (Ver Anexos) que el filtrado colaborativo recomendaría libros relacionados al que un usuario seleccione dentro de una lista de resultados de una búsqueda.

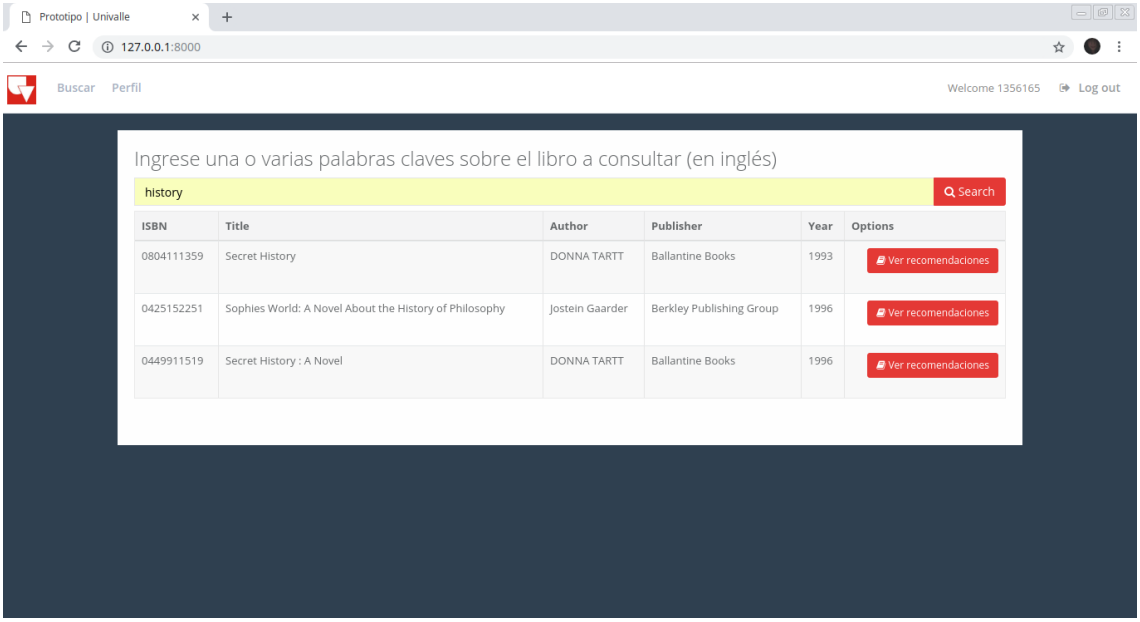

Figura 6.5: Resultados de una búsqueda

Aunque la base de datos externa que se utilizó para este módulo es mucho más pequeña que la proporcionada por la biblioteca Mario Carvajal, se seguían presentado problemas de memoria en el equipo de cómputo por lo que se realizó un filtro para trabajar con una "versión reducida" de la base de datos. Este filtro consiste en seleccionar solamente los libros que cuenten con más de 50 ratings por parte de los usuarios, de esta manera también se pueden asegurar resultados más precisos. Además se seleccionan únicamente los ratings de los usuarios cuya localización sea "usa" y "canadá" y posteriormente se eliminan registros duplicados.

6.1.5. Técnicas de recomendación de filtrado colaborativo: descomposición en valores singulares y k vecinos más cercanos

Para la recomendación por filtrado colaborativo se implementaron dos técnicas populares que hacen uso de los datos de todas las tablas que conforman la base de datos externa.

El algoritmo de la técnica SVD se puede describir en los siguientes pasos:

1. Con la matriz reducida se crea una matriz que relaciona los usuarios en las columnas y los libros

en las filas, y cuyos valores son los ratings que los usuarios le han proporcionado a cada uno de los libros con el uso de la función "pivot" del paquete pandas que manipula los dataframe.

2. Haciendo uso del paquete sklearn, se elige una función que implementa la técnica SVD para ser aplicada en la matriz user-libro y se aplica la reducción correspondiente que determina los factores latentes más importantes obteniendo como resultado una matriz aproximada y de menor dimensión.

3. A la matriz resultante del punto anterior, se aplica la función del paquete numpy llamada "corrcoef" que calcula los coeficientes de correlación de Pearson de cada uno de los pares de libros contenidos en la matriz. Una explicación más detallada de esta función se encuentra en el capítulo "Modelo" de este proyecto.

4. Para seleccionar los libros similares a cualquier libro objetivo, se buscan los coeficientes cuyos valores estén entre 0.8 y 1.0 iterando la matriz obtenida en el paso anterior.

5. Finalmente se seleccionan los n valores más altos y se retorna una lista con la información de los libros elegidos para la recomendación.

Por otro lado, la técnica de k vecinos más cercanos puede enumerarse así:

1. Tratando de optimizar el manejo de la información, inicialmente se crea una matriz comprimida utilizando una función del paquete scipy que trabaja especialmente con matrices muy dispersas y las comprime para que los procesos de iteración sean más eficientes.

2. El paquete de sklearn cuenta también con la implementación del método "vecinos cercanos" que agrupa los elementos similares del conjunto inicial de entrenamiento de acuerdo a un atributo o variable, y que al ingresar un elemento nuevo o uno del conjunto de prueba, calcula la distancia de éste con sus vecinos para determinar a qué clase o "vecindario" pertenecerá este elemento.

3. Esta función de vecinos cercanos cuenta con diferentes opciones para calcular la distancia entre elementos, usando en esta oportunidad la medida de similaridad coseno del ángulo explicada en el capítulo "Modelo".

4. El paso siguiente es definir el número k de vecindarios que se quieren obtener con la función "kneighbors" que para el proyecto se ha determinado en 10 que será el número de libros que se quiere recomendar.

5. La función retorna las distancias entre los elementos y los ids de los mismos que se utilizan para obtener toda la información correspondiente que se mostrará finalmente al usuario.

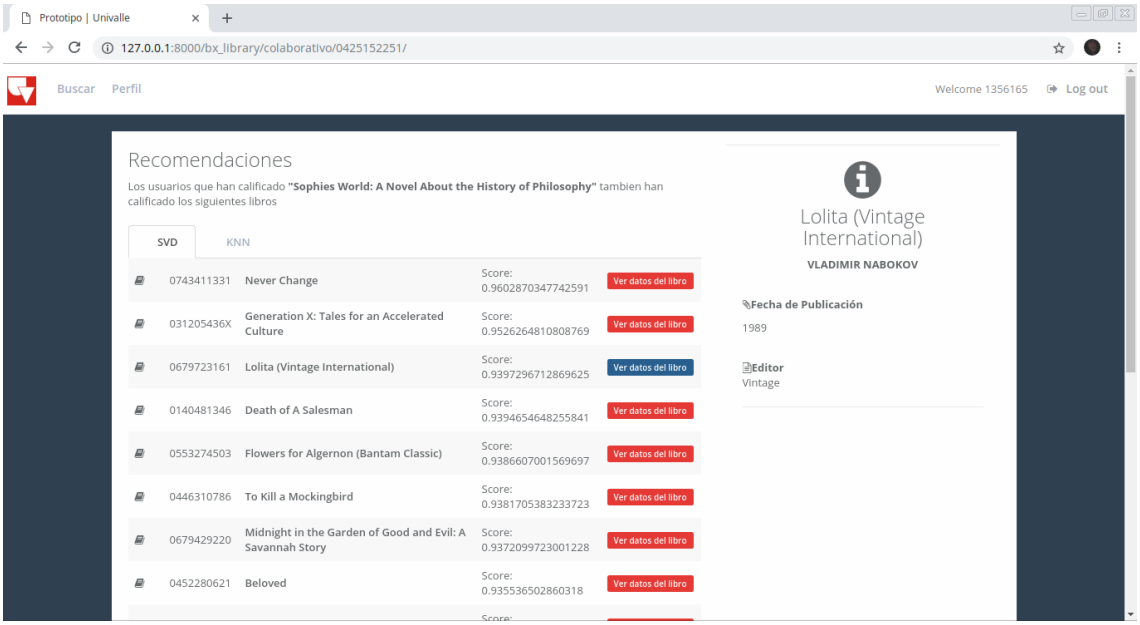

Figura 6.6: Recomendaciones basadas en filtrado colaborativo

Un diagrama de flujo con la representación de las actividades y pasos que puede seguir un usuario al hacer uso del prototipo web y obtener recomendaciones puede encontrarse en los anexos de este trabajo.

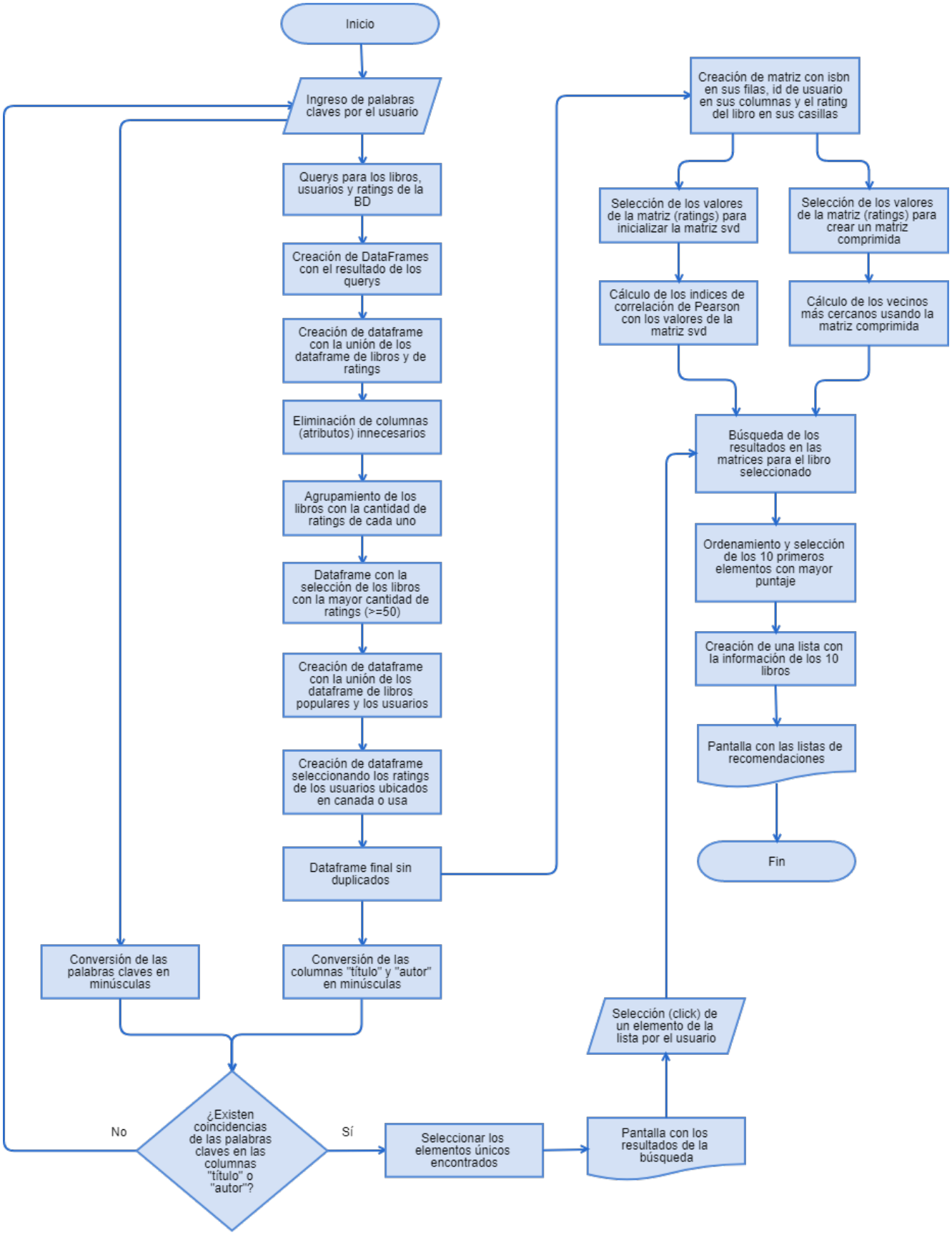

Figura 6.7: Diagrama de flujo para las recomendaciones por filtrado colaborativo

Capítulo 7

Pruebas

Teniendo en cuenta la naturaleza de un proyecto de recomendación se decide realizar dos tipos de prueba para medir la usabilidad y la precisión.

7.1. Pruebas de usabilidad

Se realizaron pruebas de usabilidad por medio de un formulario de Google donde se recibe la realimentación por parte de usuarios (estudiantes de la Universidad del Valle que han hecho uso de los servicios bibliotecarios) de los distintos aspectos de la aplicación teniendo en cuenta factores como la interactividad, la calidad de las recomendaciones y la interfaz de usuario, entre otros. Los resultados de las encuestas pueden encontrarse en el anexo.

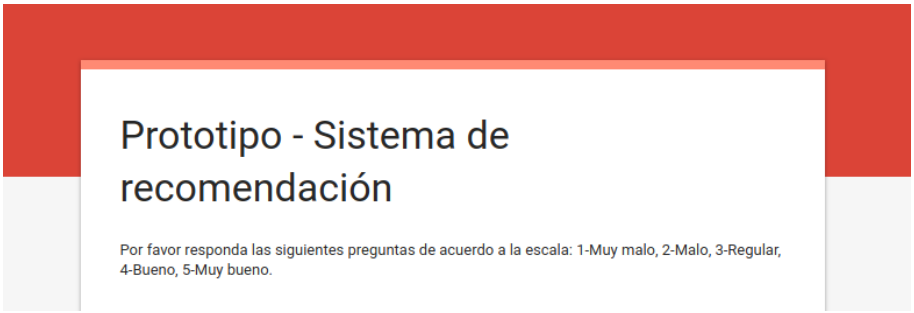

Figura 7.1: Muestra de la encuesta realizada a usuarios

El promedio para cada respuesta fue:

Recomendación de contenido	**Relevancia basado en contenido**	**Calidad de resultados de búsqueda**	**Recomendación filtrado colaborativo**	**Interfaz de usuario**	**Usabilidad**
4,130434783	4,217391304	4,326086957	4,217391304	4,173913043	4,282608696

7.2. Pruebas de precisión

Se deciden realizar pruebas automatizadas seleccionando 200 títulos de libros de la base de datos aleatoriamente, para ser usados como entrada de los algoritmos y obtener las nueve mejores recomendaciones.

Una vez se tiene la matriz de libros-scores de recomendaciones, se calcula el promedio de la enésima mejor recomendación como medida de precisión.

La razón para realizar las pruebas de esta manera se debe a que la salida de todos los algoritmos están centrados en el coeficiente de correlación de Pearson y la similaridad coseno, cuyos valores siempre se encuentran entre 0 y 1.

A diferencia de otras técnicas de recomendación donde se poseen valores de ratings para los elementos, y las pruebas que realizan comúnmente se enfocan en la comparación del valor real del rating y de la predicción hecha por el algoritmo, este enfoque no se pudo aplicar en el prototipo evitando pruebas de precision tradicionales como la matriz de confusión, entre otras.

El resultado de los experimentos realizados pueden ser encontrados en el anexo.

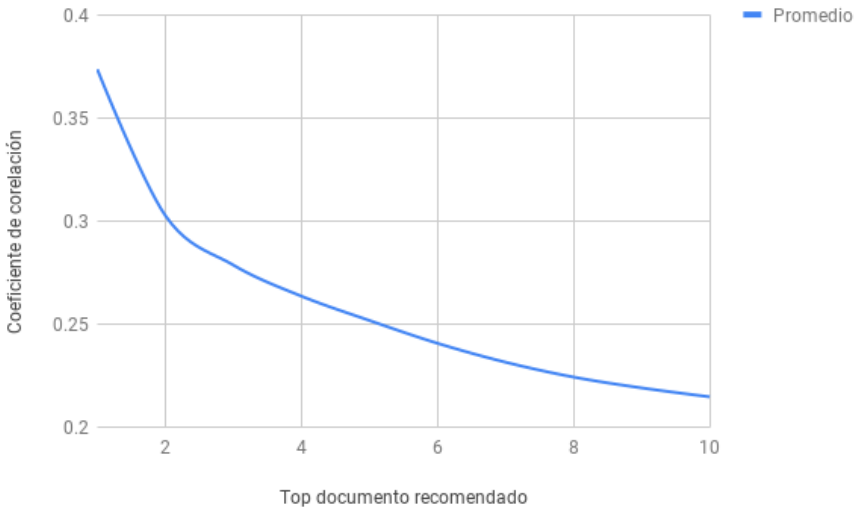

Figura 7.2: Prueba para la recomendación basada en contenido: similaridad de coseno

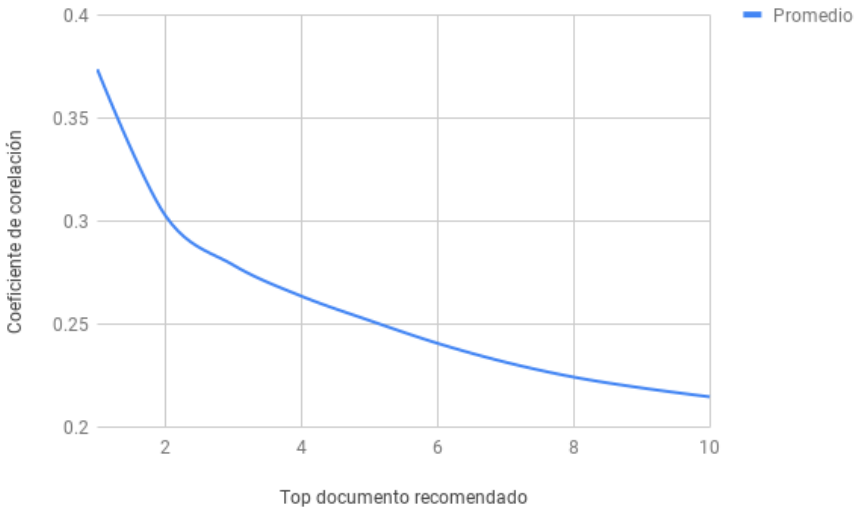

Figura 7.3: Prueba para la recomendación basada en contenido: linear kernel

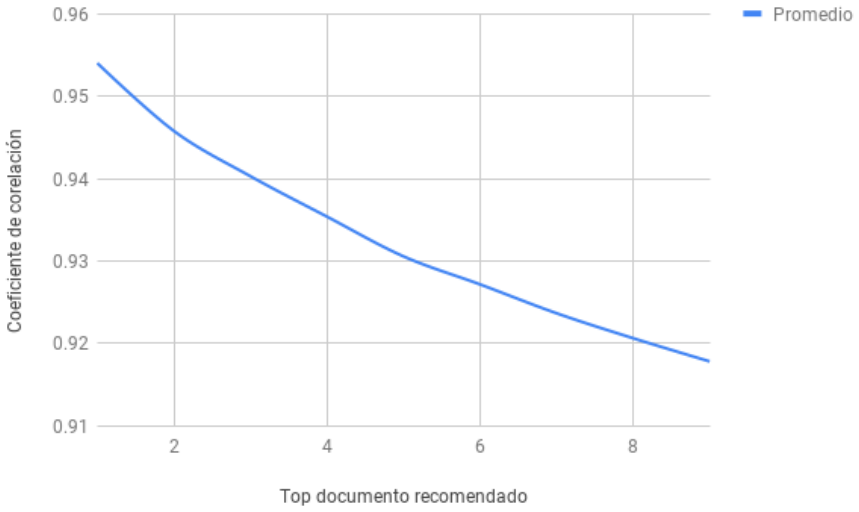

Figura 7.4: Prueba para la recomendación basada en filtrado colaborativo: descomposición en valores singulares

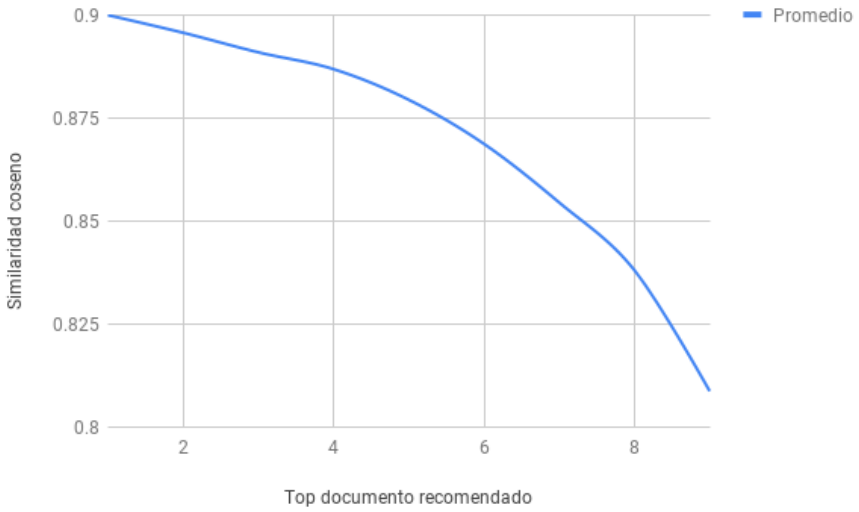

Figura 7.5: Prueba para la recomendación basada en filtrado colaborativo: vecinos más cercanos

7.3. Análisis de resultados

7.3.1. Resultados de pruebas de usabilidad

De acuerdo a los resultados promedios obtenidos en las encuestas realizadas a los usuarios se recibe un feedback positivo en general. Los usuarios manifiestan conformidad con la aplicación y relevancia en las recomendaciones obtenidas.

Las encuestas realizadas en este trabajo de grado fueron desarrolladas por estudiantes de ingeniería de sistemas de distintos semestres. Se considera importante realizar un proceso similar con estudiantes de otras carreras, profesor y distinto personal que haga uso de los servicios bibliotecarios para recibir un feedback más objetivo.

La realimentación de menor puntuación fueron los resultados obtenidos por la técnica basada en contenido siendo consecuente con la calidad de los resultados encontrados en las pruebas.

7.3.2. Resultados de las recomendaciones basadas en contenido

Con los valores obtenidos después de realizar las pruebas de precisión, se puede observar como las dos técnicas (similaridad coseno y linear kernel) muestran resultados casi idénticos como se esperaba y se explicó en los capítulos de desarrollo y modelo: al utilizar la similaridad coseno se calcula el coseno del ángulo entre los vectores, con vectores normalizados esta función se reduce al producto punto entre los vectores, que es precisamente la fórmula que utiliza la estrategia de linear kernel. La mejor recomendacion tuvo un valor promedio de 0.4 y para la última recomendacion se obtuvo un valor promedio de 0.1, valores similares a los encontrados en el estado del arte.

Hay que tener presente que dado el gran volumen de datos con el que cuenta la biblioteca Mario Carvajal, no es posible realizar un sistema de recomendación basado en contenido por la mala escalabilidad de esta técnica. Debido a esto para el desarrollo se optó por reducir el tamaño de la base de datos, creando una nueva con menos registros de títulos y transacciones. Esta reducción se llevó a cabo de manera arbitraria donde sólo se buscó disminuir el tamaño de la base de datos. Esta decisión de diseño es la adaptación que se encontró más adecuada para este problema que ya se acerca al big data donde quizá se puedan encontrar mejores aproximaciones.

Para estar seguros de que la disminución realizada no haya impactado negativamente al algoritmo, es posible que sean necesarios análisis estadísticos exhaustivos y combinatorios que se salen del alcance de este proyecto.

7.3.3. Resultados de las recomendaciones basadas en filtrado colaborativo

Por otra parte, en los resultados obtenidos por las pruebas de las técnicas de filtrado colaborativo se observan valores mucho más altos (en comparación con los resultados basados en contenido), muy cercanos al 1.0 ideal en ambas estrategias. La estrategia con mejor rendimiento fue la descomposición en valores singulares lo cual es consecuente con lo encontrado en el estado del arte y con el problema particular, dado a que esta estrategia sobresale en ámbitos donde los elementos poseen características similares.

Para las pruebas realizadas fue necesario aplicar un filtro a la base de datos completa de este caso de estudio. Este filtro consistió en seleccionar los elementos calificados por usuarios localizados en una misma región, puntualmente en transacciones de usuarios de Estados Unidos y Canadá. Se decidió restringir la búsqueda de transacciones de estos usuarios debido a que representan el mayor porcentaje dentro del total de las transacciones y tenía un tamaño adecuado para que el tiempo de ejecución del algoritmo fuera aceptable

Si bien los resultados son muy prometedores en el filtrado colaborativo, hay que resaltar también que tiene problemas serios de escalabilidad por lo que debe considerarse que en un sistema con una magnitud de usuarios como la de la biblioteca Mario Carvajal, al implementar ratings como se propone en este trabajo, llegará un momento en que será necesario aplicar filtros de este tipo para que los usuarios obtengan recomendaciones en tiempos aceptables.

Como análisis final de este trabajo de grado se concluye que deben buscarse filtros para restringir las consultas en el sistema o buscar un enfoque orientado a big data para la implementación de un sistema de recomendacion adecuado.

Capítulo 8

Conclusiones

- La formación académica recibida en la universidad me permitió hacerle frente a problemas y obstáculos encontrados durante el desarrollo del proyecto, así como también pude aprovechar todo el conocimiento obtenido durante la carrera para culminar el trabajo de grado.
- El uso del framework de aplicaciones web Django fue una gran elección para la implementación del prototipo al ser una herramienta que permite el desarrollo ágil de proyectos de pequeña a grande escala. Su principal ventaja es la facilidad de manejo incluso para desarrolladores con poca experiencia que combinado con características como su rapidez, su seguridad y escalabilidad, lo convierten en una herramienta poderosa para cualquier tipo de desarrollo web.
- Las librerías de Python numpy, pandas, sklearn y scipy son herramientas fundamentales para cualquier persona que esté llevando a cabo un proyecto relacionado con la ciencia de datos, machine learning o big data. Estas permiten, de manera relativamente sencilla, el manejo y manipulación de grandes cantidades de datos, así como la obtención de nueva información sobre los mismos.
- Aunque es posible obtener buenos resultados con las herramientas mencionadas anteriormente, en el caso concreto de la aplicación de un algoritmo de recomendación basado en contenido para la biblioteca Mario Carvajal no se pudo llegar a la misma conclusión pues se tiene una gran suma de datos a los que no se les puede dar un buen manejo con este método. Esto es por el alto costo computacional que representa la técnica basada en contenido, que toma como referencia la descripción de los libros o material bibliográfico y la magnitud de la base de datos es considerablemente superior a lo que se puede manejar con equipos de cómputo estándar.
- El método de filtrado colaborativo mostró mejores resultados como era esperado, lo que también pudo deberse al uso de una base de datos más pequeña y a la aplicación de filtros a la hora de obtener la información que conformaría la matriz inicial.
- Teniendo en cuenta los comentarios anteriores, se llega a la conclusión que una buena manera para optimizar algoritmos simples de recomendación que usan grandes cantidades de datos es el uso de alguna estrategia de filtrado de la información que reduzca el rango de elementos que se incluyen en la matriz inicial.
- Observando los resultados obtenidos en las pruebas con usuarios y las pruebas funcionales, se concluye finalmente que las técnicas de filtrado colaborativo presentan valores más altos respecto a la similaridad entre los elementos recomendados y el elemento objetivo.

Capítulo 9

Trabajos futuros

- En el alcance del proyecto se definió que el resultado de este trabajo sería un prototipo de sistema de recomendación para la biblioteca Mario Carvajal. De esta manera, el siguiente paso sería la integración de este proyecto al sistema OPAC que utiliza la biblioteca para la interacción online con sus usuarios.
- En las conclusiones se encontraron diferentes opciones que aumentan la eficiencia y precisión de los resultados de los algoritmos de recomendación, en el futuro podrían agregarse a los algoritmos diversos filtros dinámicos que se adapten a las necesidades de los usuarios y a la información que se pueda obtener de ellos, como por ejemplo la localización de la mayoría de sus transacciones con la red de bibliotecas de la Universidad del Valle o la carrera a la que pertenece.
- Otra buena alternativa para buscar mejores resultados sería probar más técnicas basadas en el contenido de la información de los libros que puedan trabajar con un rango más amplio de datos pero que no aumente considerablemente el costo computacional de la ejecución del mismo.
- Si se encuentra una manera de implementar calificaciones de los usuarios a los libros en el sistema de la biblioteca Mario Carvajal, o de medir las interacciones de los usuarios con el material bibliográfico, también sería posible probar más técnicas de filtrado colaborativo y observar si se encuentran mejores resultados en comparación con las que ya se han implementado.
- Teniendo en cuenta el volumen de información que maneja la biblioteca Mario Carvajal y que éste crece a medida que pasa el tiempo, una mejor opción que podría implementarse a futuro sería usar un enfoque basado en big data que permita un mejor manejo de la totalidad de la información y tenga en consideración otros atributos que describen el material bibliográfico de la universidad.

Capítulo 10

Bibliografía

[1] S. Russell, P. Norvig, J. C. Rodríguez, and L. J. Aguilar, *Inteligencia artificial 2nd ed.* Pearson Educacion, 2011.

[2] T. Segaran, *Programming collective intelligence. 1st ed.* O'Reilly.n, 2007.

[3] E. País, "Tres universidades del valle, entre las 20 mejores del país." `http://www.elpais.com.co/colombia/tres-universidades-del-valle-entre-las-20-mejores-del-pais.html`, 2015. Consultado 1 Abril de 2017.

[4] "Informe de gestión." `http://biblioteca.univalle.edu.co/informacion-general/documentos/informes-de-gestion`, 2015. Consultado 1 Abril de 2017.

[5] Registrounico.bibliotecanacional.gov.co., "Perfiles y funciones del bibliotecario público — biblioteca nacional de colombia." `http://registrounico.bibliotecanacional.gov.co/content/perfiles-y-funciones-del-bibliotecario-p%C3%BAblico`, 2017. Consultado 18 Abril de 2017.

[6] K. Tsuji, E. Kuroo, S. Sato, U. Ikeuchi, A. Ikeuchi, F. Yoshikane, and H. Itsumura, "Use of library loan records for book recommendation," *IIAI International Conference On Advanced Applied Informatics*, 2012.

[7] K. Tsuji, F. Yoshikane, S. Sato, and H. Itsumura, "Book recommendation using machine learning methods based on library loan records and bibliographic information," *IIAI 3Rd International Conference On Advanced Applied Informatics*, 2014.

[8] L. Xin, E. Haihong, S. Junde, S. Meina, and T. Junjie, "Collaborative book recommendation based on readers' borrowing records.," *International Conference On Advanced Cloud And Big Data*, 2013.

[9] S. Kanetkar, A. Nayak, S. Swamy, and G. Bhatia, "Web-based personalized hybrid book recommendation system," *International Conference On Advances In Engineering & Technology Research*, 2014.

[10] S. Parvatikar and B. Joshi, "Online book recommendation system by using collaborative filtering and association mining," *IEEE International Conference On Computational Intelligence And Computing Research (ICCIC)*, 2015.

[11] P. Mathew, B. Kuriakose, and V. Hegde, "Book recommendation system through content based and collaborative filtering method," *International Conference On Data Mining And Advanced Computing (SAPIENCE)*, 2016.

[12] C. Chen, L. Zhang, H. Qiao, S. Wang, Y. Liu, and X. Qiu, "Book recommendation based on book-loan logs," *The Outreach Of Digital Libraries: A Globalized Resource Network*, 2012.

[13] F. Jia and Y. Shi, "Library management system based on recommendation system," *Communications In Computer And Information Science*, 2013.

[14] J. A. Konstan and J. Riedl, "Recommender systems: From algorithms to user experience," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1-2, pp. 101–123, 2012.

[15] P. Symeonidis, D. Ntempos, and Y. Manolopoulos, "Recommender Systems for Location-based Social Networks," 2014.

[16] G. Kembellec, G. Chartron, and I. Saleh, *Recommender systems*. 2014.

[17] L. Kidzinski and H. S. Nguyen, "Statistical foundations of recommender systems," no. September, 2011.

[18] H. Ren and W. Feng, "CONCERT: A concept-centric web news recommendation system," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7923 LNCS, pp. 796–798, 2013.

[19] Y. Gao, W. Yu, P. Chao, R. Zhang, A. Zhou, and X. Yang, "A Restaurant Recommendation System by Analyzing Ratings and Aspects in Reviews," in *Database Systems for Advanced Applications* (M. Renz, C. Shahabi, X. Zhou, and M. A. Cheema, eds.), (Cham), pp. 526–530, Springer International Publishing, 2015.

[20] S. Uslu, C. Ozturan, and M. F. Uslu, "Course scheduler and recommendation system for students," *Application of Information and Communication Technologies, AICT 2016 - Conference Proceedings*, 2017.

[21] H.-C. Chen and A. L. P. Chen, "A music recommendation system based on music data grouping and user interests," *Proceedings of the tenth international conference on Information and knowledge management - CIKM'01*, p. 231, 2001.

[22] A. John, H. Muhammed Ilyas, and V. Vasudevan, "Medication recommendation system based on clinical documents," *Proceedings - 2016 International Conference on Information Science, ICIS 2016*, pp. 180–184, 2017.

[23] L. Song, C. Wang, X. Duan, B. Xiao, X. Liu, R. Zhang, X. He, and X. Gong, "Taxihailer: A situation-specific taxi pick-up points recommendation system," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8422 LNCS, no. PART 2, pp. 523–526, 2014.

[24] E. Elisabeth, R. Nock, and F. Célimène, "Demonstrator of a tourist recommendation system," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8302 LNCS, pp. 171–175, 2013.

[25] T. Alashkar, S. Jiang, and Y. Fu, "Rule-Based Facial Makeup Recommendation System," *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heterogeneous Face Recognition, HFR 2017,*

Joint Challenge on Dominant and Complementary Emotion Recognition Using Micro Emotion Features and Head-Pose Estimation, DCER and HPE 2017 and 3rd Facial Expression Recognition and Analysis Challenge, FERA 2017, pp. 325–330, 2017.

[26] C. B. Pop, V. R. Chifu, I. Salomie, A. Stetco, and R. Plaian, "Lifestyle Recommendation System for Treating Malnutrition," in *Intelligent Distributed Computing VIII* (D. Camacho, L. Braubach, S. Venticinque, and C. Badica, eds.), (Cham), pp. 41–46, Springer International Publishing, 2015.

[27] K. Bhumichitr, S. Channarukul, N. Saejiem, R. Jiamthapthaksin, and K. Nongpong, "Recommender Systems for university elective course recommendation," *Proceedings of the 2017 14th International Joint Conference on Computer Science and Software Engineering, JCSSE 2017*, 2017.

[28] R. Kawase, B. P. Nunes, and P. Siehndel, "Content-based movie recommendation within learning contexts," *Proceedings - 2013 IEEE 13th International Conference on Advanced Learning Technologies, ICALT 2013*, pp. 171–173, 2013.

[29] K.-H. N. Bui, X. H. Pham, J. J. Jung, O.-J. Lee, and M.-S. Hong, "Context-Based Traffic Recommendation System," in *Context-Aware Systems and Applications* (P. C. Vinh and V. Alagar, eds.), (Cham), pp. 122–131, Springer International Publishing, 2016.

[30] K. Zhang, Z. Zhang, K. Bian, J. Xu, and J. Gao, "A Personalized Next-Song Recommendation System Using Community Detection and Markov Model," *Proceedings - 2017 IEEE 2nd International Conference on Data Science in Cyberspace, DSC 2017*, pp. 118–123, 2017.

[31] J. H. Kim, C. W. Song, K. W. Lim, and J. H. Lee, "Design of music recommendation system using context information," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4088 LNAI, pp. 708–713, 2006.

[32] S. K. Ko, S. M. Choi, H. S. Eom, J. W. Cha, H. Cho, L. Kim, and Y. S. Han, "A smart movie recommendation system," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6771 LNCS, no. PART 1, pp. 558–566, 2011.

[33] X. Shi, W. Mi, L. Huang, Y. Niu, D. Chang, and Y. Zhang, "An Intelligent Recommendation System for TV Programme," in *2012 International Conference on Information Technology and Management Science(ICITMS 2012) Proceedings* (B. Xu, ed.), (Berlin, Heidelberg), pp. 109–118, Springer Berlin Heidelberg, 2013.

[34] J. Nessel and B. Cimpa, "The MovieOracle - Content based movie recommendations," *Proceedings - 2011 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2011*, vol. 3, pp. 361–364, 2011.

[35] H. Li, F. Cai, and Z. Liao, "Content-based filtering recommendation algorithm using HMM," *Proceedings - 4th International Conference on Computational and Information Sciences, ICCIS 2012*, pp. 275–277, 2012.

[36] S. N. Ferdous and M. M. Ali, "A semantic content based recommendation system for cross-lingual news," *2017 IEEE International Conference on Imaging, Vision and Pattern Recognition, icIVPR 2017*, 2017.

[37] F. Zhang, "A Personalized Time-Sequence-Based Book Recommendation Algorithm for Digital Libraries," *IEEE Access*, vol. 4, pp. 2714–2720, 2016.

[38] N. Kurmashov, K. Latuta, and A. Nussipbekov, "Online book recommendation system," *Proceedings of the 2015 12th International Conference on Electronics Computer and Computation, ICECCO 2015*, pp. 3–6, 2016.

[39] P. Devika, R. C. Jisha, and G. P. Sajeev, "A novel approach for book recommendation systems," *2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016*, 2017.

[40] S. Kanetkar, A. Nayak, S. Swamy, and G. Bhatia, "Web-based personalized hybrid book recommendation system," *2014 International Conference on Advances in Engineering and Technology Research, ICAETR 2014*, pp. 0–4, 2014.

[41] S.-Y. Hwang and E.-P. Lim, "A Data Mining Approach to New Library Book Recommendations *," *Lncs*, vol. 2555, pp. 229–240, 2002.

[42] Y. Teng, L. Zhang, Y. Tian, and X. Li, "A novel FAHP based book recommendation method by fusing apriori rule mining," *Proceedings - The 2015 10th International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2015*, pp. 237–243, 2016.

[43] T. Desai, S. Gandhi, P. Murlidhar, S. Gupta, M. Vijayalakshmi, and G. P. Bhole, "An enterprise-friendly book recommendation system for very sparse data," *International Conference on Computing, Analytics and Security Trends, CAST 2016*, pp. 211–215, 2017.

[44] A. S. Tewari and K. Priyanka, "Book recommendation system based on collaborative filtering and association rule mining for college students," *Proceedings of 2014 International Conference on Contemporary Computing and Informatics, IC3I 2014*, pp. 135–138, 2014.

[45] P. Jomsri, "Book recommendation system for digital library based on user profiles by using association rule," *4th International Conference on Innovative Computing Technology, INTECH 2014 and 3rd International Conference on Future Generation Communication Technologies, FGCT 2014*, pp. 130–134, 2014.

[46] A. S. Tewari and A. G. Barman, "Collaborative book recommendation system using trust based social network and association rule mining," *Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016*, no. 2, pp. 85–88, 2016.

[47] S.-H. Sie and J.-H. Yeh, "Library Book Recommendations Based on Latent Topic Aggregation," *Proceedings of 16th International Conference on Asia-Pacific Digtal Libraries*, pp. 411–416, 2014.

[48] R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization," in *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, (New York, NY, USA), pp. 195–204, ACM, 2000.

[49] G. Semeraro, P. Basile, M. De Gemmis, and P. Lops, "Content-based recommendation services for personalized digital libraries," *Lecture Notes in Computer Science including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, vol. 4877 LNCS, pp. 77–86, 2007.

[50] Wikipedia, "Norma vectorial." `http://https://es.wikipedia.org/wiki/Norma_vectorial`. Consultado 19 agosto de 2018.

[51] I. de Monterey, "Logaritmo de un cociente." `https://www.montereyinstitute.org/courses/DevelopmentalMath/TEXTGROUP-15-19_RESOURCE/U18_L2_T2_text_final_es.html`. Consultado 19 agosto de 2018.

[52] C.-N. Ziegler, "Book-crossing dataset." `http://www2.informatik.uni-freiburg.de/~cziegler/BX/`, 2004. Consultado 19 agosto de 2018.

[53] "pandas - dataframe." `http://pandas.pydata.org/pandas-docs/version/0.23.4/generated/pandas.DataFrame.html`, 2018. Consultado 19 agosto de 2018.

[54] "sklearn - tfidfvectorizer." `http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`, 2018. Consultado 19 agosto de 2018.

Anexos

Modelo entidad relación

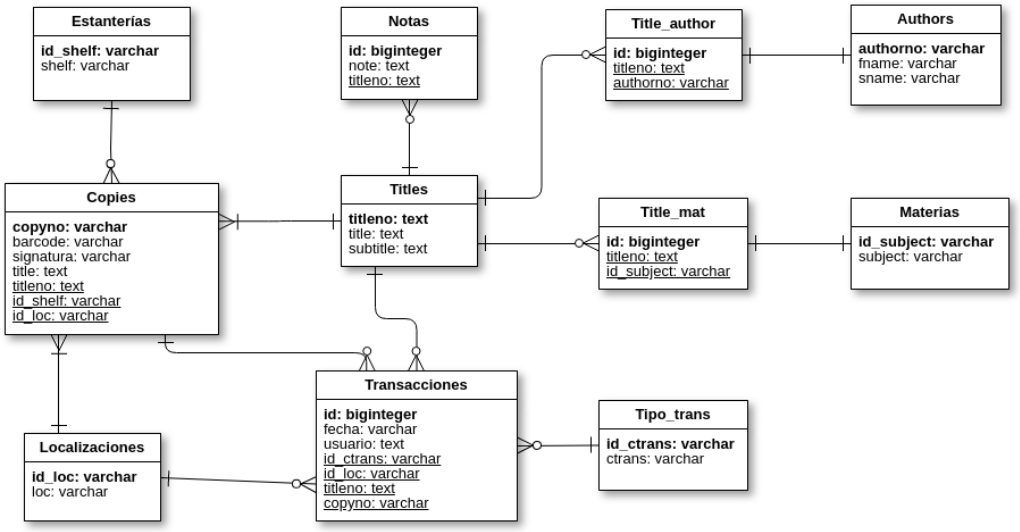

Figura 10.1: Modelo entidad relación de la base de datos de la biblioteca Mario Carvajal

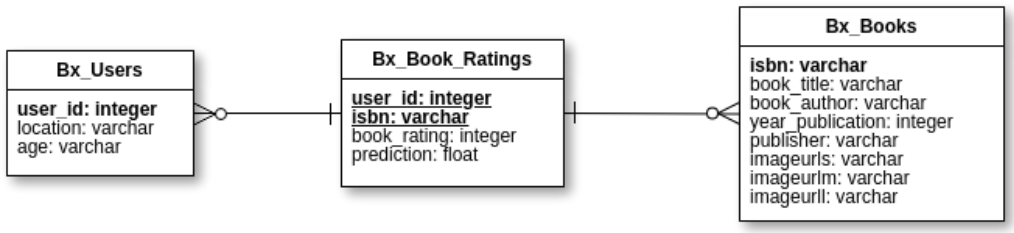

Figura 10.2: Modelo entidad relación de la base de datos externa

Diagrama de Despliegue

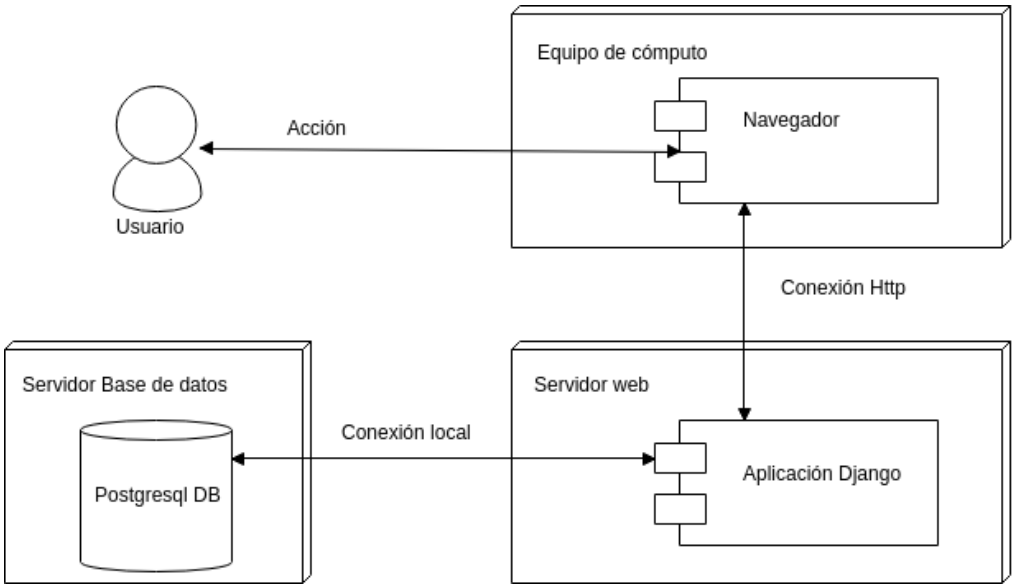

Figura 10.3: Diagrama de despliegue del proyecto

Diagrama de flujo

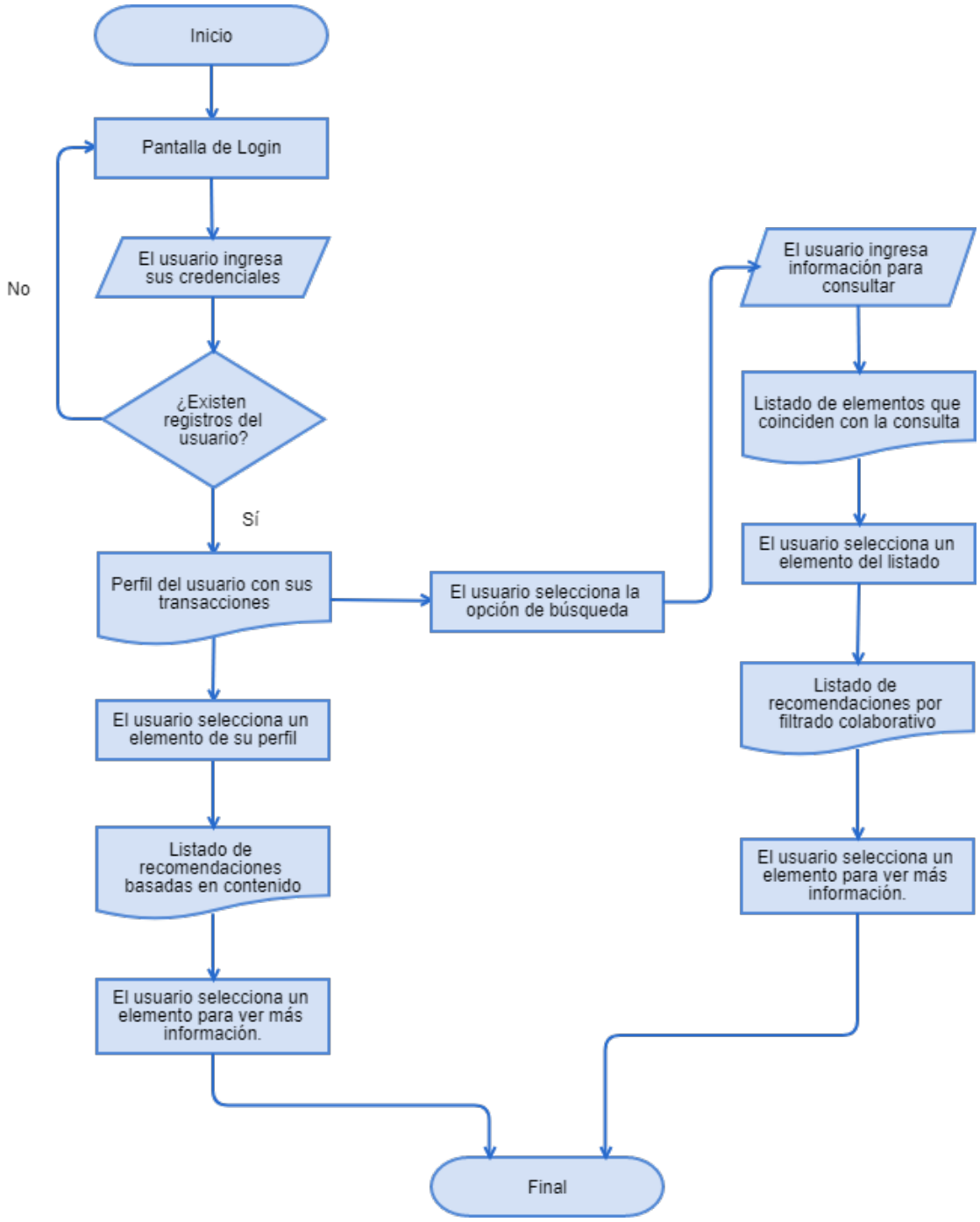

Figura 10.4: Diagrama de flujo para un usuario