Proyecto de Sistema de Recomendación de Filtrado Colaborativo basado en Machine Learning

Cesar Enrique Pita Perez
Postgrado en Informática
Universidad Mayor de San Andrés
La Paz – Bolivia
cesar.aapv@gmail.com

***Resumen*—En el presente trabajo se describe conceptos de sistemas de recomendación, machine learning y la aplicación de estas tecnologías en diferentes áreas. Se realiza un estudio del estado del arte sobre los sistemas de recomendación y se presenta el diseñó de la investigación.**

Palabras clave—Sistema de recomendación, filtrado colaborativo, machine learning.

I. INTRODUCCIÓN

Debido al avance tecnológico hoy en día las empresas tanto públicos como privados generan una gran cantidad de información relativa a sus productos, contenidos y/o servicios que en algunos casos son expuestos mediante sus sistemas de información [1].

La información que es expuesta a los usuarios y/o clientes por parte de estas empresas contiene en muchos casos una infinidad de productos, servicios y/o contenidos cada uno con características diferentes que fueron diseñados para llenar las expectativas de cada uno de sus clientes o usuarios. Ante esta variedad de productos, contenidos y servicios los usuarios intentan discriminar la información presentada y seleccionar un producto que cubra sus necesidades, gustos y/o preferencias de manera óptima.

A. Planteamiento del Problema

Muchas de las empresas en Bolivia tanto públicas como privadas, ante el inminente avance tecnológico invierten en el desarrollo de portales web de tal forma que puedan presentar y brindar sus bienes y/o servicios en línea a sus clientes y/o usuarios. Si bien invierten recursos económicos, humanos y tiempo en el desarrollo de páginas web o aplicaciones móviles, éstos no generan valor o ventaja competitiva si no se enfocan en las necesidades o preferencias de sus clientes y/o usuarios.

Las empresas son conscientes de la importancia de la fidelización de los clientes y/o usuarios para el rendimiento de la empresa, en este sentido un cliente y/o usuario satisfecho probablemente volverá a comprar o recurrir a sus servicios. En tal sentido una buena recomendación es sinónimo de fidelización del cliente. Esta sugerencia debe satisfacer las necesidades del cliente; esto implica conocer bien los gustos y preferencias de éste.

Con la llegada del comercio electrónico y la compra de productos, contenidos y/o servicios por internet es imprescindible que las páginas web desarrolladas cuenten con algún mecanismo que les permita conocer los gustos y preferencias de sus clientes y/o usuarios.

Los sistemas de recomendación surgen como una respuesta a estas necesidades, estos sistemas recopilan datos que son ingresados por los usuarios; datos que son capturas de manera explícita con algún mecanismo de puntuación a cada ítem; y datos que son capturados de manera explícita en base a la actividad registrada por el usuario. En base a este historial de datos es que se realizan predicciones para sugerir artículos.

Dado que las expectativas de los usuarios siempre están cambiando y que estos modelos generalmente se basan en modelos basados en memoria, surge el problema de la calidad de la predicción que es ofrecida al usuario, ¿Cómo lograr mayor efectividad en la predicción de preferencias de usuarios?

Los sistemas de recomendación aplicando técnicas de Machine Learning son una respuesta para lograr una mayor efectividad en las recomendaciones y de esta manera lograr que las organizaciones a través de sus portales web puedan llegar al mayor número de clientes y/o usuarios, con el fin de elevar sus ingresos o brindar un mejor servicio.

B. Formulación del Problema de Investigación

¿En qué grado mejorara las métricas de predicción de preferencias de usuario si se utiliza un modelo basado en Machine Learning?

C. Planteamiento de Objetivos

1) Objetivo General

Plantear las fases para el desarrollo de un modelo de Sistema de Recomendación de Filtrado Colaborativo utilizando técnicas de Machine Learning que permita mejorar las métricas de predicción de preferencias de usuarios.

2) Objetivos Específicos

- Diagnosticar sistemas de recomendación automáticas.
- Identificar vulnerabilidades y puntos a mejorar.
- Diseñar un mecanismo de recuperación de datos que sea eficaz y fácil de implementar.
- Identificar las fases metodológicas para el diseño del Modelo de Machine Learning para el sistema de recomendación.
- Proponer la secuencia de pasos necesarios para el desarrollo del Sistema de Recomendación.

Para referenciar este artículo (IEEE):

[N] C. Pita, «Proyecto de Sistema de Recomendación de Filtrado Colaborativo basado en Machine Learning», *Revista PGI. Investigación, Ciencia y Tecnología en Informática*, nº 8, pp. 48-51, 2020.

D. Planteamiento de Hipótesis

La elaboración de un modelo de Sistema de Recomendaciones basado en Machine Learning permitirá mejorar las métricas de predicción de preferencias de usuarios.

TABLA I. OPERACIONALIZACIÓN DE VARIABLES

Variables	Componentes	Dimensiones	Indicadores
Métricas de predicción de preferencias de usuarios.	- Analizar la experiencia del Cliente y/o usuario.	Mecanismos para efectividad y/o fiabilidad de la predicción	Indicador KPI CRR (Customer Retention Rate).
			- RSME (Error cuadrático medio)
Modelo de sistema de recomendaciones basado en machine learning.	Capacidad del modelo para inducir datos.	Eficacia del modelo.	- MAE Error del modelo al realizar predicciones.

E. Marco Teórico

1) Sistemas de Recomendación

Los sistemas de recomendación forman parte de un sistema de filtrado de información, los cuales presentan distintos tipos de temas o ítems de información (películas, música, libros, noticias, imágenes, páginas web, etc.) que son del interés de un usuario en particular. Generalmente, un sistema recomendador compara el perfil del usuario con algunas características de referencia de los temas, y busca predecir el "ranking" o ponderación que el usuario le daría a un ítem que aún el sistema no ha considerado. Estas características pueden basarse en la relación o acercamiento del usuario con el tema o en el ambiente social del mismo usuario [9].

2) Recomendaciones Basadas en Contenido

Los sistemas basados en contenido realizan recomendaciones utilizando tanto las características del usuario como de los ítems que le han interesado al usuario [14].

Gran parte de la investigación de este tipo de sistemas de recomendación se ha enfocado en recomendar ítems con contenido textual, tales como páginas web, libros y películas. Por eso, muchas aproximaciones han tratado este problema como una tarea de recuperación de información, donde el contenido asociado con las preferencias del usuario es tratado como una consulta y los ítems no calificados son puntuados por su relevancia para dicha consulta [4].

3) Filtrado Colaborativo

Los sistemas basados en filtrado colaborativo, utilizan solamente la información contenida en la matriz de utilidades para realizar recomendaciones y sus modelos se basan en las similitudes que presentan las calificaciones entre pares de usuarios o de ítems [1].

Fig. 1. Filtrado Colaborativo. Portal inmobiliario.com

4) Aproximaciones Híbridas

Este tipo de filtrado combina las dos estrategias anteriores. El problema de recomendar ítems que no sean parecidos a los que el usuario suele calificar positivamente puede ser solucionado con el filtrado colaborativo, ya que, basándonos en el comportamiento pasado de otros usuarios podemos encontrar ítems novedosos para el usuario actual. El problema del arranque en frío que tienen los sistemas de filtrado colaborativo puede ser solucionado parcialmente aplicando las técnicas basadas en contenido utilizando, por ejemplo, la información demográfica del usuario para encontrar los ítems calificados por usuarios similares [13].

5) Machine Learning

Machine Learning (técnicas de aprendizaje automático) es un conjunto de métodos capaces de detectar patrones en un conjunto de datos para realizar predicciones o para otro tipo de toma de decisiones [5].

F. Aprendizaje Supervisado y No Supervisado

El aprendizaje supervisado es cuando se tiene una variable de salida, sea cualitativa o cuantitativa que se desea predecir en base a un conjunto de características. Se debe establecer un modelo que relacione el conjunto de características y la variable de salida. Para el aprendizaje supervisado se debe considerar un conjunto de datos de entrenamiento que permite predecir los valores de salida [12].

Fig. 2. Aprendizaje supervisado. [5]

G. Algoritmos de Aprendizaje Supervisados

1) K vecinos más cercanos (Hearest-neightbor)

El método de machine learning de los K vecinos más cercanos (kNN - k nearest neighbor) es uno de los más simples y generalmente se refiere a un aprendizaje perezoso ya que el aprendizaje no se implementa realmente hasta que la clasificación o la predicción es requerida [7].

2) Naive Bayes

El clasificador de tipo Naive Bayes, es un modelo probabilístico eficiente basado en el teorema de Bayes, el cual examina la probabilidad de que características aparezcan en las clases predichas [11].

3) Tabla de Decisión

Un clasificador de tipo tabla de decisión se construye sobre la idea conceptual de una tabla de búsqueda. El clasificador retorna la clase mayoría del conjunto de entrenamiento si la celda de la tabla de decisión que coincide con la nueva instancia está vacía. En ciertos conjuntos de datos, se puede conseguir una mayor performance de clasificación usando tablas de decisión en vez de otros modelos más complejos [3].

4) Máquinas de Soporte Vectorial

Son capaces de manejar de manera eficiente los datos multidimensionales. Originalmente fueron diseñadas como un clasificador de dos clases, aunque pueden funcionar con más clases realizando múltiples clasificaciones binarias (una a una

entre cada par de clases). El algoritmo funciona clasificando instancias basadas en una función lineal de la característica [6].

5) Redes Neuronales Artificiales

Una red neuronal artificial (*ANN - artificial neural network*) es un grupo interconectado de nodos con la intención de representar la red de neuronas en el cerebro [3].

6) Arboles de decisión

Los árboles de decisión son clasificadores muy utilizados debido a que el algoritmo crea reglas que son fáciles de entender e interpretar [3].

7) Ensambles

Un ensamble (conjunto) es una colección de múltiples clasificadores base que toman un nuevo ejemplo, pasado a cada uno de los clasificadores base, y luego combina esas Predicciones de acuerdo a algún método, como por ejemplo a través del voto. La motivación es que, mediante la combinación de las predicciones, el conjunto es menos probable de clasificar erróneamente [3].

H. Estado del Arte

1) Recomendaciones en tiempo real mediante filtrado colaborativo incremental y real time big data

En el presente trabajo se desarrolla una plataforma trabajo bajo filtrado colaborativo planteando un algoritmo que se pueda actualizar de forma incremental en memoria, implementando una arquitectura de respuesta en tiempo real. Se realizó el diseño de la lógica de la aplicación y la interfaz de usuario. Dentro los resultados obtenidos pudieron comprobar que el algoritmo desarrollado satisface el objetivo planteado y que es capaz de aprender incrementalmente [2].

2) Inducción de preferencias a partir del contexto de elección del usuario en sistemas de recomendación

En este trabajo de tesis presentamos las modificaciones hechas a los algoritmos clásicos de filtrado colaborativo basado en memoria para que utilicen el contexto de elección del usuario al momento de predecir sus preferencias por nuevos ítems. Como no se conocen conjuntos de datos públicos que tengan el contexto de elección del usuario y que permitan probar los algoritmos modificados, implementamos dichas modificaciones en un sistema de recomendación real para poder recolectar los datos necesarios para la etapa de experimentación. Finalmente, se realizó una serie de experimentos sobre tres conjuntos de datos, que permitieron verificar que la propuesta tiene un mejor desempeño que los sistemas de recomendación clásicos [6].

3) Clasificación multilingüe de documentos utilizando machine learning y la Wikipedia

En el trabajo realizado se propone la utilización de machine learning para la clasificación automática de documentos, utilizando para este fin se utiliza una representación de modelo bolsa (Bag of Words, BoW) basada en Wikipedia. La principal contribución es la elaboración de un modelo para la clasificación monolingüe y multilingüe de los documentos de texto.

Al final del documento se puede apreciar que los resultados obtenidos con el modelo propuesto son bastante ventajosos y que si corroboran la hipótesis planteada [8].

II. MÉTODOS

A. Marco Metodológico

1) Diseño Metodológico

Para el desarrollo del presente trabajo de investigación es realizada por medio de una investigación tecnológica no experimental en las ciencias de la computación que presenta una serie de características que la vinculan a la innovación tecnológica. Con innovación tecnológica se designa la incorporación de nuevos procesos orientados al cliente o consumidor final. Para nuestro caso en la incorporación de sistemas de recomendación automáticas [10].

2) Tipo de Investigación

Se utilizará un tipo de investigación descriptiva en el que se examinará el problema en cuestión y se describirá los componentes que comprenden el modelo propuesto.

3) Método de Investigación

El método de investigación será deductivo pues se realizará una investigación de principios generales de sistemas de recomendación para deducir por medio del razonamiento lógico los aspectos que involucran la aplicación de un modelo de este tipo.

4) Fases Metodológicas

La metodología que se seguirá durante el desarrollo de investigación se basará en las siguientes fases:

Fase 1: Recopilación y clasificación de información
Fase 2: Planteamiento del problema.
Fase 3: Definición de Hipótesis y Objetivos.
Fase 4: Desarrollo de la investigación.
Fase 5: Conclusiones

B. Técnicas de Investigación

1) Observación. En base la experiencia de trabajo, se podrá realizar las observaciones de los flujos o actividades o procesos de la entidad para la implementación de un sistema de recomendaciones.

2) Test. Para obtener datos sobre los usuarios y/o clientes

C. Universo o población de referencia

El presente trabajo está dirigido a los portales encargados de difundir películas en formato digital.

D. Delimitación geográfica

La muestra o población de estudio está enfocada a la información recolectada de los usuarios que calificaron películas en MovieLens.

E. Delimitación temporal

Los datos que serán considerados para la realización del trabajo de investigación propuesto serán enmarcados en base al dataset publicado por Movielens (2018).

F. Método de Investigación

Dentro los sistemas de recomendación de filtrado colaborativo existen dos clasificaciones, basados en memoria y basado en modelos.

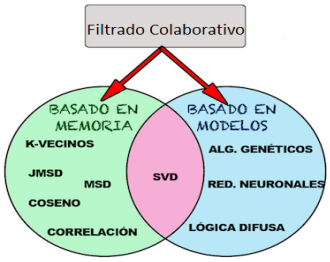

Fig. 3. Filtrado Colaborativo. [4]

La gran mayoría de los trabajos estudiados trabajan sobre modelos basados en memoria, los más utilizados K-NN y Correlación. Estos modelos no son complejos de implementar, pero presentan problemas cuando de escalamiento y procesamiento cuando el volumen de los datos va creciendo.

SVD es un método que utiliza alguna de las características de los sistemas basados en modelos y los aplica a los basados en memorias. En los trabajos revisados se puede apreciar que en las técnicas SVD se mejoran las métricas de predicción.

Los métodos basados en modelos no se implementaron debido a la complejidad que representa el procesamiento de los datos y el diseño del modelo.

Hoy en día este problema ya no es una limitante debido al gran avance tecnológico. En la revisión bibliográfica, la gran mayoría indica que el volumen de datos es un factor importante para optimizar las métricas de predicción.

Machine Learning, es una tecnología ampliamente utilizada en diferentes áreas del conocimiento con excelentes resultados. Agrupar conjuntos de datos con determinadas características, predecir comportamientos e identificar patrones son algunas de las tareas que podemos realizar con esta tecnología.

Considerando estos avances tecnológicos, se plantea diseñar un sistema de recomendación de filtrado colaborativo basado en machine learning.

III. Resultados esperados

Como planteamiento de solución se trabajará en una solución hibrida en el ámbito del filtrado colaborativo.

Para el problema de arranque en frio se propone una clasificación de los videos más populares. Se tiene planificado trabajar con un dataset proporcionado por MovieLens, el cual contiene la calificación de películas por parte de los usuarios.

A continuación de describe la secuencia de pasos como propuesta de solución:

1. Identificación de variables.
2. Tratamiento de datos categóricos.
3. División de dataset en conjunto de entrenamiento y pruebas
4. Escalado de variables para agilizar los resultados.
5. Construir la Red Neuronal Artificial.
6. Definir la capa de entrada y las capas ocultas.
7. Ajustar la red neuronal al conjunto de entrenamiento
8. Evaluar el modelo y calcular predicciones.
9. Validación cruzada.

IV. Conclusiones

Los sistemas de recomendación han cobrado gran importancia en organizaciones que ofrecen productos/servicios, ya que incrementan sus utilidades y fidelizan a sus clientes.

Desde sus inicios estos sistemas han evolucionado con el fin de mejorar sus métricas de predicción.

En el presente trabajo se presenta una propuesta de modificación al algoritmo de filtrado colaborativo aplicando redes neuronales con el fin de mejorar las métricas de predicción.

Referencias

[1] C. A. R. Morales, «Algoritmo SVD aplicado a los sistemas de recomendación en el comercio,» *Tecnología, Investigación y Academia*, p. 10, 2018.

[2] P. P. Yague, P. R. Jiménez y M. M. Patiño, «Recomendaciones en tiempo real mediante filtrado colaborativo incremental y rel time Big Data,» Madrid, 2019.

[3] D. P. Sastre, Deep Learning para Sistemas de Recomendación basados en contenido, Madrid: Propia, 2017.

[4] C. C. Aggarwal, Recommender Systems, New York: Springer, 2016.

[5] J. C. P. Gallegos, S. A. Torres y F. S. Quezada Aguilera, Inteligencia Artificial, Europa Aid.: Iniciativa Latinoamericana de Libros de Texto Abiertos., 2014.

[6] A. Roberto, Inducción de Preferencias a partir del contexto de elección del usuario en sistemas de recomendación, Buenos Aires, Argentina.: Propia, 2013.

[7] J. Ávila, Sistemas de Recomendaciones de Contenido de TV digital basado en Ontologías., Cuenca, Ecuador: Propia., 2014.

[8] M. A. M. García, Clasificación multilingüe de documentos utilizando machine learning y la Wikipedia, Vigo, España.: EIDO, 2017.

[9] M. C. M. Álvaro y F. J. Segovia Perez, Sistema de Recomendación híbrido para la predicción de calificaciones em Yelp.com, Madrid, España: U. Politécnica, 2016.

[10] C. M. A. Álvarez, Metodología de la investigación cuantitativa y cualitativa, Neiva: Universidad Sur Colombiana, 2011.

[11] E. R. N. Valdez, Sistemas de Recomendación de Contenidos para Libros Inteligentes, Oviedo: Propia, 2012.

[12] R. Alvarado, J. Hernández y E. Villatoro, «Sistema de recomendación de música basado en aprendizaje semi supervisado,» *Research in Computing Science 94*, p. 13, 2015.

[13] D. Jannach, M. Zanker y A. Feldering, Recommender Systems an Introductions, Estados Unidos: Cambridge, 2011.

[14] V. Patricia, C. Cornelis y M. De Cock, Trust Networks for Recommender Systems, Ámsterdam Paris: Atlantis Press, 2011.

Breve CV del autor

Cesar Enrique Pita Perez es Ingeniero de Sistemas por la Universidad Técnica de Oruro. Diplomado en Planificación y Gerencia de Sistemas. Actualmente realiza la maestría en Alta Gerencia y en TICs e Innovación para el Desarrollo MAG-TIC en el Postgrado en Informática de la Universidad Mayor de San Andrés. Inscrito a la SIB con Matrícula 28122
Profesional en desarrollo en el Ministerio de Economía y Finanzas Públicas. Sus intereses: Investigación de Operaciones, Ciencia de Datos, Ingeniería de Datos. Email: cesar.aapv@gmail.com.