Propuesta de un conjunto de datos sintético para estudiar el problema de arranque en frío en sistemas de recomendaciones

Julio Herce-Zelaya
Dept. Ciencias de la Computación e I.A.
Universidad de Granada
Granada, España
julioherce@gmail.com

Daniel Ranchal-Parrado
Dept. Ciencias de la Computación e I.A.
Universidad de Granada
Granada, España
danielranchal@correo.ugr.es

Carlos Porcel
Dept. Ciencias de la Computación e I.A., DaSCI
Universidad de Granada
Granada, España
cporcel@decsai.ugr.es

Álvaro Tejeda-Lorente
Dept. Ciencias de la Computación e I.A.
Universidad de Granada
Granada, España
atejeda@decsai.ugr.es

Juan Bernabé-Moreno
Dept. Ciencias de la Computación e I.A.
Universidad de Granada
Granada, España
jbernabemoreno@gmail.com

Enrique Herrera-Viedma
Dept. Ciencias de la Computación e I.A., DaSCI
Universidad de Granada
Granada, España
viedma@decsai.ugr.es

***Abstract*—Los sistemas de recomendaciones ayudan en la selección de ítems relevantes para ser sugeridos a los usuarios. Uno de los desafíos más importantes al que se enfrentan es el conocido como problema de arranque en frío, que ocurre cuando se incorporan al sistema nuevos usuarios o ítems de los que no se tiene información previa. Este trabajo plantea un método de obtención de datos provenientes de diversas fuentes sociales e introduce un conjunto de datos sintético que ayude a estudiar y mitigar este problema, aportando información sobre usuarios, películas y valoraciones. Además, con la idea de probar y trabajar con dicho conjunto de datos se propone un esquema híbrido para generar recomendaciones precisas en situaciones de arranque en frío. Los resultados validan la eficacia del conjunto de datos presentado de cara a futuras investigaciones en sistemas de recomendaciones, especialmente en contextos de arranque en frío.**

***Index Terms*—Sistemas de recomendaciones, problema de arranque en frío, conjunto de datos sintético, enfoque híbrido.**

I. Introducción

Los sistemas de recomendaciones son herramientas que ayudan a los usuarios en el proceso de toma de decisiones a la hora de elegir ítems que puedan serles relevantes y ajustados a sus preferencias, entre una gran cantidad de opciones disponibles. Estos sistemas utilizan información previa de los usuarios para generar recomendaciones personalizadas. Esta información la pueden suministrar los usuarios de manera explícita o bien se puede obtener de forma implícita observando su comportamiento. Una vez obtenidos los perfiles de los usuarios y caracterizados los ítems a recomendar, puede empezar a funcionar el enfoque adoptado para generar las recomendaciones. Por ejemplo, se pueden usar esquemas basados en contenido, evaluando ítems de características similares a los valorados anteriormente por el usuario, esquemas colaborativos, recomendando ítems valorados positivamente por usuarios con un perfil similar o bien, esquemas híbridos que combinan elementos de diversos enfoques de recomendación.

Sin embargo, cuando un usuario es nuevo en el sistema, no se disponen de datos previos y, por lo tanto, no es posible generar recomendaciones personalizadas para ese usuario. Esta situación es lo que se denomina problema de arranque en frío. Los métodos más comunes para paliar este problema consisten en consultar a los usuarios sobre sus preferencias o pedirles directamente que valoren algunos ítems y así tener algunos datos con los que poder empezar a generar recomendaciones. El principal inconveniente de estos métodos es que requieren tiempo y esfuerzo por parte de los usuarios por lo que en muchos casos se ignoran y termina afectando tanto a la calidad de las recomendaciones como a las experiencias con el sistema.

En este trabajo se presenta un resumen del artículo titulado ”Introducing CSP Dataset: A Dataset Optimized for the Study of the Cold Start Problem in Recommender Systems” [1] en el que se propone un método para obtener información

Esta publicación es parte del proyecto de I+D+i PID2022-139297OB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER/UE

Partially supported by Spanish Thematic Network on Recommender Systems (Action RED2022-134302-T funded by MCIN/AEI/10.13039/501100011033)

implícita de los usuarios a partir de los datos que insertan en sus redes sociales, lo que permite generar sus perfiles sin necesidad de interacción explícita. Este método se usa para construir un nuevo conjunto de datos optimizado para abordar el problema de arranque en frío. En concreto, se compone de tres tablas sobre películas, perfiles de usuario y valoraciones, obtenidas a partir de datos recopilados de Filmaffinity y Twitter. Este conjunto de datos es especialmente valioso para crear y estudiar esquemas de recomendaciones con situaciones de arranque en frío.

A diferencia de otros conjuntos de datos disponibles, el propuesto ofrece hasta 12 variables relacionadas con el comportamiento del usuario, permitiendo modelos de decisión más potentes y precisos que aprovechan esos datos obtenidos de manera implícita. Se describe el uso del conjunto de datos realizando un proceso de limpieza del mismo, selección de características destacables y, por último, diseño y evaluación de un esquema de recomendaciones híbrido que combina filtrado colaborativo y basado en contenido, con resultados favorables.

II. Materiales y métodos

El conjunto de datos propuesto se extrajo de las redes sociales Filmaffinity y Twitter, e incluye datos de películas, valoraciones y perfiles de usuarios. Desde Filmaffinity se extrajeron datos de películas y valoraciones, mientras que desde Twitter se extrajeron los perfiles de los usuarios. La extracción se realizó mediante una técnica de *web scraping*, usando las páginas de perfil de Filmaffinity donde los usuarios pueden vincular sus cuentas de Twitter, permitiendo así la correlación entre los datos de valoraciones y los perfiles de los usuarios.

El conjunto de datos obtenido se aloja en Github [1] (acceso el 19 de marzo de 2024) y consta de tres tablas en formato CSV: películas, valoraciones y perfiles de usuario. Cada tabla contiene información detallada, como títulos de películas, valoraciones y datos específicos de los usuarios de Twitter, como el número de seguidores o la hora preferida para enviar sus mensajes y comentarios.

Para la importación, limpieza y procesamiento de los datos, se utilizó el lenguaje Python con las bibliotecas *pandas* y *NumPy*, entre otras. El proceso de limpieza de datos requirió la agregación y transformación de algunos campos numéricos en datos de tipo categórico y la eliminación de campos irrelevantes. Además, se generaron nuevas características a partir de los datos existentes, mejorando así la información representada y ayudando al entrenamiento de los posibles modelos.

Por último, el esquema híbrido usado para la generación de recomendaciones combina un enfoque basado en contenidos, que clasifica las características de los usuarios y se asignan a las de los ítems, con un filtrado colaborativo en el que se consideran las valoraciones de usuarios con perfiles similares.

[1] https://github.com/lynchblue/movie-rating-dataset

III. Resultados

Para evaluar el esquema planteado con el conjunto de datos obtenido, se procedió a su división usando el 80% para entrenamiento y el 20% para test. Esta división se realizó en función de los ID de los usuarios con la idea de que los usuarios pertenecientes a los datos de test no aparezcan en los datos de entrenamiento, emulando así el escenario de arranque en frío de nuevos usuarios.

De la lista de resultados se eligieron los 5 primeros y para la evaluación del modelo se usaron las métricas de precisión, rango recíproco medio y media de ítems recomendados. También se realizó un estudio comparativo con otro esquema sobre recomendaciones para completar conjuntos de ítems. Los resultados obtenidos muestran la bondad del método propuesto para el tratamiento del problema de arranque en frío facilitando la formación de perfiles de usuario y, por tanto, mejorando la precisión del sistema sin afectar a las experiencias de los usuarios.

IV. Conclusiones

En el trabajo se propone un conjunto de datos que ofrece tablas que reflejan tanto el comportamiento de los usuarios como sus valoraciones de películas y una tabla con las características de las películas. Se describe el método que permite obtener de forma implícita el conjunto de datos a través de información extraída de las redes sociales de los usuarios (Filmaffinity y Twitter). De esta forma se facilita la caracterización de los usuarios y se pueden generar predicciones precisas al correlacionar datos de su comportamiento con sus valoraciones.

Este conjunto de datos emerge como un recurso valioso por la escasez de conjuntos de datos que proporcionen características sobre los ítems y comportamiento de usuarios, pudiendo usarse como un estándar para investigar el problema de arranque en frío. Los resultados de los experimentos respaldan la hipótesis de que el conjunto de datos presentado se puede utilizar para crear recomendaciones para los usuarios sin tener información previa de ellos, abriendo líneas para futuras investigaciones que mejoren la precisión de las recomendaciones y exploren nuevas características a partir de los datos disponibles.

La limitación principal del trabajo radica en que para aplicar el método planteado y los modelos basados en el conjunto de datos, los usuarios necesitan tener una cuenta en Twitter o similar para poder obtener sus perfiles de comportamiento. Como trabajo futuro, se pueden estudiar otras técnicas para mejorar la precisión de las recomendaciones, adaptar el método de adquisición de preferencias y completarlo con nuevas características a partir de los datos sin procesar.

References

[1] J. Herce-Zelaya, C. Porcel, Á. Tejeda-Lorente, J. Bernabé-Moreno, and E. Herrera-Viedma, 'Introducing CSP Dataset: A Dataset Optimized for the Study of the Cold Start Problem in Recommender Systems', Information, vol. 14, no. 1, 2023.