Informe del proyecto de sistema de recuperacion de información

David Sánchez Iglesias

Universidad de La Habana, Facultad de Matemática y Computación

En los últimos años, con el desarrollo de internet y el aumento de la cantidad de personas que utilizan la red, se ha incrementado también enormemente la cantidad de información que se encuentra disponible en la red. Como resultado, surgieron algoritmos y técnicas para recuperar información de manera eficiente. Muchos de estos están destinados a mejorar la experiencia de los usuarios al buscar o incluso comprar en línea. Debido a la cantidad masiva de usuarios que un sitio web de ventas puede tener diariamente conectados a la vez, y a la cantidad de información que cada uno genera y la cantidad de productos que cada uno puede potencialmente vender y comprar, ofrecer los mejores productos para cada cliente de manera que estos satisfagan sus necesidades y se amolden a sus gustos se ha convertido en un problema de gran importancia. Para poder ofrecer dichos productos a las personas correctas es necesario extraer y organizar cierta cantidad de datos de los usuarios y de los productos, así como usarlos de manera inteligente. Así, aparecen las reseñas, información que los usuarios dan a los administradores del sitio para saber cómo se sintieron con su compra, por qué se sintieron así y por qué aprueban el producto o no. En este proyecto se propone un sistema de recuperación de información que use estos datos y genere información útil para que los vendedores puedan ofrecer los productos correctos a los clientes correctos, maximizando así sus ganancias y la probabilidad de que el usuario comprador obtenga el producto que busca de la manera más rápida y eficiente posible.

1. Porblema de ejemplo

Se tiene un sitio web de ventas en línea así como las reseñas de un gran número de usuarios acerca de los productos que este sitio web vende. Se quiere saber cuáles productos recomendar a cada usuario de manera que a este le parezca más atractivo y lo estimule a comprarlo.

2. Estado del arte

En la actualidad existen muchos sistemas de recomendación que se basan en la información de los usuarios y de los productos para ofrecer productos a los usuarios. Algunos de estos sistemas son:

- Collaborative filtering: Este método se basa en la idea de que si a un usuario le gustan ciertos productos, entonces le gustarán también los productos que

a otros usuarios que tienen gustos similares les gustan. Este método se basa en la información de los usuarios y de los productos.
- Content-based filtering: Este método se basa en la idea de que si a un usuario le gustan ciertos productos, entonces le gustarán también los productos que sean similares a estos. Este método se basa en la información de los productos.
- Hybrid methods: Estos métodos combinan los dos anteriores para ofrecer recomendaciones a los usuarios.

Plataformas como Amazon, Netflix y Spotify usan estos métodos para ofrecer productos a sus usuarios y, debido a la cantidad masiva de usuarios que tienen diariamente y que reseñan sus productos, la utilización de esas reseãs puede ser de gran ayuda para mejorar la calidad de las recomendaciones que se les ofrecen a los usuarios. En este proyecto se propone un sistema de recuperación que extrae y organiza dicha información. Es importante aclarar que este sistema no se encarga de ofrecer recomendaciones a los usuarios, sino de extraer y organizar la información de las reseñas para que los vendedores puedan ofrecer mejores propuestas de ventas a los usuarios.

3. Descripción del sistema

Para la realización de este proyecto se usaron rese´ nas de Amazon sobre productos digitales (programas informáticos) que recibieron al menos una reseña por parte de algún usuario (no se usó todo el dataset por su inmenso tamaño: casi 50 gb de reseñas, de ahí la necesidad de reducir el tamaño, y por eso se escogió solamente la muestra relativa a programas digitales). Lo primero que se suele buscar y que suele llamar la atención de los vendedores son los productos más populares, los más vendidos. Independientemente de si gustaron o no por la naturaleza de sus reseñas, estos productos responden a las necesidades de la gente o son del agrado de la mayoría, de ahí que sea importante saber cuáles son. El método para extraerlos es bastante sencillo: se crea una lista de todos los productos vendidos y se ordena de mayor a menor según la cantidad de veces que fue comprado por una persona. El dataset de Amazon ofrece, para cada reseña, si el usuario que la escribió compró o no el producto. De esta manera, se puede saber cuántas veces fue comprado un producto por la cantidad de reseñas que tiene. Independientemente de si son populares o no, aquellos productos que despiertan la curiosidad de los usuarios, que son más polémicos, que generan más reseñas, e incluso los que no, los menos polémicos, son también de interés para los vendedores. Para ellos, se hace una busqueda enfocada en los productos. Por cada producto se cuenta la cantidad de reseñas que tiene y la lista de productos resultante se ordena de mayor a menor según la cantidad de reseñas que tiene. Otra estrategia que se enfoca en los productos es usar la clasificación por estrellas. Puede haber quien no desea o no puede en el momento redactar una reseña. En esos casos, la clasificación por estrellas es de gran ayuda. Se cuenta la cantidad de reseñas que tiene cada producto y se ordena de mayor a menor según la cantidad de estrellas que tiene. E incluso si sí tiene una reseña

escrita, la clasificación por estrellas es de gran ayuda para saber si el producto es bueno o no ya que constituye la forma más rápida de clasificar un producto para un usuario segun su utilidad y/o calidad. Y es por eso que este proyecto también extrae tanto los productos con mejor clasificación de estrellas como los de menor. Luego, saliendo ya de las estrategias enfocadas en los productos, se suele concentrar la atención en los usuarios y el tipo de producto que compran. Sabiendo qué productos ha comprado una persona, sería posible predecir cuáles otros de los que no ha comprado es más probable que necesite o que se vea estimulado a comprar. Para ello, la estrategia que se suele usar es comparar el registro de compras, los productos que ha comprado, con el de otros usuarios. De este modo, se puede intuir que el primer usuario tiene "las mismas necesidades o gustos"que aquellos usuarios que han comprado lo mismo. Así, es posible que si dos personas han hecho las mismas compras, entonces es probable que compren los mismos productos en el futuro. Para extraer estos productos, se compara el registro de compras de cada usuario con el de los demás y se extraen aquellos productos que los usuarios con mayor similitud en sus compras han comprado y que el usuario en cuestión aún no. Este método es bastante eficaz cuando el número de usuarios en la plataforma es grande y cuando ya la persona en cuestión posee un historial de compras mas o menos extenso. Mientras más productos haya comprado, más predecible puede resultar su siguiente compra usando este método. Por último, en este proyecto se propone el uso de los textos de las propias reseñas. Usando un método de .ªnálisis de sentimiento"se puede detectar si el texto de una reseña es positivo o negativo. Esto puede informar mucho acerca de la calidad del producto en cuestión. Por ejemplo, es posible que un producto sea un éxito de ventas pero no porque sea bueno o de gran calidad, sino porque responde a una necesidad general de la gente hasta el momento no satisfecha o porque al menos, una de sus características es satisface dicha necesidad. En este caso, si el producto posee muchas reseñas egativas pero es muy vendido, se puede intuir cuál es la necesidad que satisface y se puede usar esta información para mejorar el producto o para ofrecer productos similares a los usuarios. Por otro lado, si el producto es muy vendido y posee muchas reseñas positivas, entonces se puede intuir que el producto es de gran calidad y que satisface las necesidades de los usuarios. En este proyecto se propone el uso de un método de análisis de sentimiento para extraer los productos con más reseñas positivas y los que tienen más reseñas negativas. Para esto último, se usaron las bibliotecas de Python *nltk* y *sklearn* para el análisis de sentimiento. Se usó el método de *Desition Tree (árbol de decisión)* para clasificar los textos de las reseñas en positivos y negativos. Para entrenar el modelo se usó el dataset *amazon polarity*, un dataset también de reseñas de Amazon que posee un fragmento para entrenamiento y uno de tests, preparado para este tipo de análisis.