Sistemas de Recomendación

Alejandra Esquivel
alejandraeg9899@gmail.com

Septiembre, 13, 2015

1. Objetivos

Este trabajo busca esbozar las características de una máquina de recomendación, su aplicación en los negocios modernos que manejan grandes volúmenes de información en sus sitios. Por otro lado identificamos los algoritmos que se utilizan en su diseño. Mostraremos cómo se obtienen las preferencias de los usuarios para un caso simplre y aplicaremos un ejemplo de caso de estudio para buscar identificar patrones en la información usando el algoritmo de vecinos cernanos.

Los sistemas de recomendación(SR) son componentes cruciales para los negocios en linea, con ellos se busca mantener a sus usuarios satisfechos e incrementar el número de ventas diarias por usuario. Por ello considero que el conocimiento de la aplicación de la mineria de datos es una de las tendencias más importantes de la computación en nuestro tiempo.

2. Introducción

Los SR aparecen a mediados los 80 de manera natural, por los vendedores que buscaban entender los que sus clientes deseaban comprar, se generaban catálogos impresos y se entregaban con cupones dentro de sus hogares o en sus buzones de correo. Con la entrada del internet y el auge del correo electrónico se inició el acercamiento usando listas de correo, pronto se vio la necesidad de integrarlos a el sitio corporativo del negocio para dar variedad al catálogo. Gracias a los avances en los algoritmos de minería de datos se observó que algunas herramientas estadísticas podían usarse para anticiparse a las preferencias si conocíamos el perfil de otros usuarios con características similares. Fue así que iniciaron como simple observación de las preferencias de los clientes y en un inicio sólo se ofrecían listas de los artículos más comprados por otros usuarios o de temporada, con frecuencia los individuos confiaban en las recomendaciones ofrecidas por otros; de la misma manera que alguien te recomienda un libro o una película. La escala de evaluación comenzó en la mayoría de los productos con una escala fija (A = bueno, B = regular, C = malo) en donde se establece el valor de cierta variable (calidad, precio, disponibilidad) es así que las recomendaciones de cine se basan en críticas de profesionales[7]. Los sistemas de recomendación juegan un rol importante en los principales sitios web (Amazon,Netflix, IMDb)[1]. La meta de ellos es incrementar las ventas y presentar experiencias de usuario personalizadas ofreciendo sugerencias para artículos desconocidos potencialmente interesantes para un usuario.

Los SR actuales nacen de la intención de imitar este comportamiento (sugerir los mejores productos) y su concepto está dividido en tres partes:

El interés en esta área permanece alto debido a que constituye un problema rico en investigación y debido a abundancia de aplicaciones prácticas que ayuden a los usuarios a lidiar con sobrecarga de información.

Las grandes compañías de medios fueron las primeras en invertir en máquinas de aprendizaje comerciales. En 2006 Netflix anunció su máquina de aprendizaje y la competencia de minería de datos con 1 millón de dólares en premio el cual fue reclamado en 2009, con toda la atención de los medios, lo que se conoció como 'Recomendaciones de Netflix: Más allí de las 5 estrellas' lo que reveló conocimiento práctico acerca de lo que realmente importa y no solo para los SR si no que para cualquier aplicación de aprendizaje máquina comercial. La meta de Netflix Prize fue fondear un algoritmo de recomendaciones que pudiera entregar 10 % de mejora en precisión de predicción sobre el sistema existente. Apple basa su sistema de recomendaciones de estrenos en el sistema de crítica Rotten Tomatoes. Google Play Store en un sistema de ranking de aplicaciones.

El rol clave de los sistemas de recomendación resulta en una vasta cantidad de investigación en este campo

Sin embargo, a pesar de todos estos avances, la actual generación de sistemas de recomendación aún requieren mejoras para realizar métodos más efectivos y aplicables a un rango amplio de casos como recomendaciones vacacionales, ciertos tipos de servicios bancarios o de financiamiento a inversionistas, y productos a ser vendidos en una tienda creada por un çarrito inteligente". Estas mejoras incluyen mejores métodos para representar comportamiento y la información acerca de los artículos ha ser adquiridos, métodos avanzados de recomendación, incorporación de información contextual y utilización de ratings multicriterio, además del desarrollo de métodos menos intrusivos que también se apoyan en

métricas para determinar desempeño de los sistemas de recomendación.

- La recolección de preferencias de los usuarios: **No tiene nada que ver con los perfiles de usuario** ya que esto se realiza a través de una encuesta que permite conocer las preferencias de los usuarios, algunos de ellos mencionan las características deseables de un artículo específico.
- Análisis: **En esta etapa se detectan patrones en las opciones seleccionadas por los usuarios. -**
- Generación de opciones: Los SR se modifican continuamente debido a que el usuario interacciona con el catalogo de artículos y el SR debe adaptarse dinámicamente a dichos cambios.
- Artículos Recomendados: Los artículos pueden ser en general cualquier bien o servicio requerido por un usuario específico. No se requiere que el usuario tenga experiencia previa con el uso del sistema principal. Sin embargo sus selecciones son tomadas en cuenta para mejorar la precisión de la recomendación próxima.

2.1. Carácteristicas clave que un SR debería cumplir

- **Incrementar el número de artículos vendidos**: Debería ser capaz de vender un conjunto de artículos de modo que puedan ser comprados sin la intervención de los SR, es decir puede tener su propia meta de venta (** ningún visitante se puede ir sin comprar **).
- **Vender artículos diversos**: Se prefiere la diversidad de artículos al ofertar productos ya que las empresas buscan que los usuarios (clientes) detecten productos en los que ni siquiera han pensado adquirir. Con frecuencia se dan descuentos o rebajas en ellos lo que ocasiona que las recomendaciones de los usuarios impacten su venta.
- **Incrementar la satisfacción del usuario**: Un SR bien diseñado cambia la interfaz de usuario según las preferencias de los mejores clientes, ofreciendo objetivos resaltados y posibilidad de que en base a los cambios de la interfase se crean grupos de interés para ofertar productos.
- **Mejor entendimiento de lo que el usuario quiere**: El sondeo adecuado de las preferencias del usuario, permite afinar los parámetros del SR con el fin de acertar en el "mejor" producto.
- **Incrementar la fidelidad del usuario**: La interacción por parte del usuario con el sitio permite que la información sea dinámica (contenido que mantenga la atención) con frecuencias las sugerencias y reseñas de un producto mantienen al usuario mas tiempo en el sitio lo que se aprovecha dando mas opciones de compra.

2.2. Clasificación de los SR

Los SR usualmente son clasificados en las siguientes categorías:

- **Recomendaciones Basadas en contenido**: Al usuario le serían recomendados artículos similares a los que selecciona en el pasado.
- **Recomendaciones Colaborativas**: Al usuario le serían recomendados artículos que gustan a las personas con preferencias y gustos similares en el pasado.
- **Aproximación Híbrida**: Estos métodos combinan métodos colaborativos y basados en contenido.

Adicionalmente los sistemas de recomendación que predicen valores absolutos de rating que usuarios individualmente no han marcado aun en artículos no conocidos, se les conoce como *filtrado basado en preferencias* .

3. Importancia de los SR

El interés en esta área permanece alto debido a que constituye un problema rico en investigación y debido a abundancia de aplicaciones prácticas que ayuden a los usuarios a lidiar con sobrecarga de información.

Las grandes compañías de medios fueron las primeras en invertir en máquinas de aprendizaje comerciales. En 2006 Netflix anunció su máquina de aprendizaje y la competencia de minería de datos con 1 millón de dólares en premio el cual fue reclamado en 2009, con toda la atención de los medios, lo que se conoció como 'Recomendaciones de Netflix: Más allí de las 5 estrellas' lo que reveló conocimiento práctico acerca de lo que realmente importa y no solo para los SR si no que para cualquier aplicación de aprendizaje máquina comercial. La meta de Netflix Prize fue fondear un algoritmo de recomendaciones que pudiera entregar 10 % de mejora en precisión de predicción sobre el sistema existente. Apple basa su sistema de recomendaciones de estrenos en el sistema de crítica Rotten Tomatoes. Google Play Store en un sistema de ranking de aplicaciones.

El rol clave de los sistemas de recomendación resulta en una vasta cantidad de investigación en este campo

Sin embargo, a pesar de todos estos avances, la actual generación de sistemas de recomendación aún requieren mejoras para realizar métodos más efectivos y aplicables a un rango amplio de casos como recomendaciones vacacionales, ciertos tipos de servicios bancarios o de financiamiento a inversionistas, y productos a ser vendidos en una tienda creada por un "carrito inteligente". Estas mejoras incluyen mejores métodos para representar comportamiento y la información acerca de los artículos ha ser adquiridos, métodos avanzados de recomendación, incorporación de información contextual y utilización de ratings multicriterio, además del desarrollo de métodos menos intrusivos que también se apoyan en métricas para determinar desempeño de los sistemas de recomendación.

4. Antecedentes

Las raíces de los sistemas de recomendación inician con trabajos en ciencia cognitiva, recuperación de información y algunas conexiones con administración científica, emergen como un área independiente a mediados de 1990 cuando los investigadores se enfocan en problemas de recomendación que explícitamente se basaban en una estructura de rating. Intuitivamente, esta estimación es usualmente basada en la escala definida por un usuario acerca de una breve información. A partir del rating de algunos artículos se puede determinar el rating de algunos que no han sido seleccionados, con el **rating superior estimado** . De manera formal el problema de recomendación puede ser formulado como sigue: Sea C el conjunto de todos los usuarios y sea S el conjunto de los posibles artículos que pueden ser recomendados tales como libros, películas o restaurantes. El espacio S de los posibles artículos puede ser muy amplio, alcanzando los cientos de millones de artículos. Similarmente el espacio del usuario puede ser bastante amplio. Sea u la función de utilidad que mide el beneficio de un articulo s al usuario. De modo que $C \times S \rightarrow R$, donde R es la totalidad de un conjunto ordenado. Entonces, para cada usuario $c \in C$, queremos seleccionar tal $s' \in S$ que maximiza la utilidad del usuario. De manera simplificada tenemos que: $\forall c \in C, s' = argmaxu(c, s)$

En un sistema de recomendación la utilidad de un artículo es usualmente representada por un *rating* el cual indica como a un usuario particular le gusta un artículo en particular. Juan Perez le dio a "Harry Potter" el rating de 7 (en escala de 1 a 10).

Ratings. Rotten Tomatoes (Tomatómetro): El rating del tomatometro se basa en las opciones publicadas por críticos de cine y televisión, es una medida confiable de la calidad de una película y representa el porcentaje de reseñas positivas dadas a una película,

Filtrado Colaborativo: La idea detrás del filtrado colaborativo es que se pueden usar los rating de los usuarios que comparten gustos similares para predecir los que aún no han sido definidos. Para obtener intuición, se comparan los ratings por pares del usuario

4.1. Ejemplos de SR:

- Airbnb. Sitio de recomendación de hospedaje.
- Yelp. Recomendación de restaurantes.
- Los SR de grandes empresas como Google Play, Apple Movies y Netflix utilizan las reseñas escritas ahí, para mejorar las sugerencias de los artículos.

5. Descripción del Problema

Consideremos una empresa dedicada a la venta, renta de inmuebles cuyo unico método de venta es a través entrevistas, o mediante llamadas telefónicas. Ha decidido invertir en un sitio web donde ahora publica el listado de propiedades a ofertar. Como en la mayoria de sitios web el diseño se centra unicamente en ser un portal informativo y poca interacción con el usuario. Por resultado los pocos clientes que visitan el sitio observan que resulta muy complicado localizar alguna propiedad relevante. O las que se muestran como relevantes están fuera de su presupuesto.

Este ejemplo es el caso de la mayoria de negocios mexicanos que utilizan un portal web para anunciarse y no aprovechan la interacción con sus clientes. Dentro de la ciudad de morelia, hemos detectado poco más de 30 inmobiliarias que utilizan portales para promoción, incluso las grandes corporaciones (Arko, Habicasa) utilizan solo como informativo su portales. Muy pocos son los que realizan alguna encuesta o registro para conocer a sus clientes.

5.1. Metodología

Nuestra propuesta de diseño se centra en 3 puntos.

1. Crear un Marco de Datos de las inmobiliarias de la ciudad de morelia.
2. Identificar usuarios potenciales que deseen adquirir propiedades.
3. Proporcionar sugerencias cercanas a las deseadas.
4. Medir la precisión a través de encuestas de satisfacción al cliente.

5.2. Recolección de los datos

Para la recolección hemos utilizado un crawler publico que visita las paginas web de las inmobiliarias, algunas requieren post - procesamiento para limpiar el html y convertirlo a un archivo csv con el cual realizaremos la clasificación de las propiedades.

5.3. Clasificación de las propiedades

Para el procesamiento de los datos decidimos utilizar el lenguaje de programación R, para analizar y clasificar. Las propiedades que descargamos de los web sites las hemos agrupado en un solo listado con aproximadamente 500 propiedades, las cuales vienen listadas por Estado (Usado, Nuevo, Construcción), Area construida (m^2), Zona (Colonia o barrio de Referencia), Precio, Latitud, Longitud y algunas otras características deseadas. El algoritmo de clasificación para esta sección que hemos seleccionado es el de vecinos cercanos ya que estamos trabajando con variables categoricas, nuestro objetivo es analizar si podemos definir clases de propiedades.

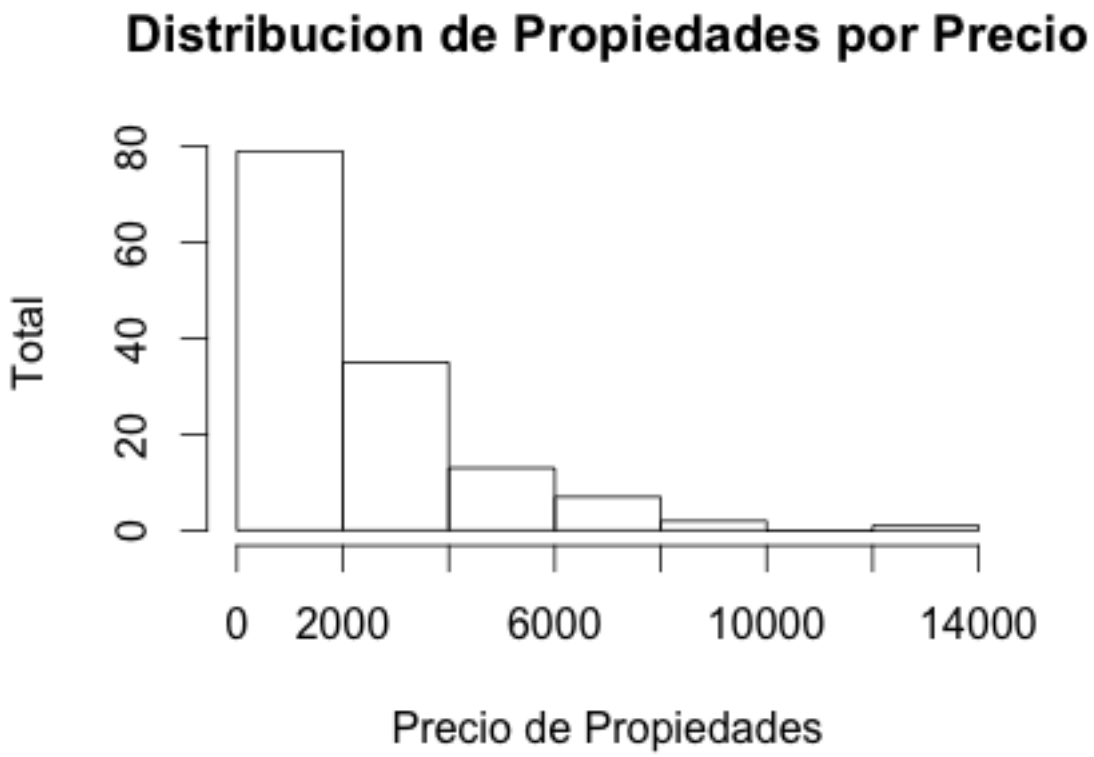

Figura 1: texto

5.4. Conclusiones

Los sistemas de recomendación han logrado una gran precisión, sin embargo pero cuando el mejor desempeño se ha alcanzado aparece la llamada *barrera mágica*[5] que se refiere a algunos niveles desconocidos de precisión, lo que revela que los ratings de los usuarios pueden ser afectados por inconsistencias en la información , básicamente *ruido*. En la mayoría de sistemas de recomendación aparecen inconsistencias y ello produce que la calidad de las recomendaciones se vea afectada.

6. Bibliografía

[1] Ricci, Francesco. Recommender Systems Handbook. New York: Springer, 2011. Print

[2] Adomavicius, G., and A. Tuzhilin. "Toward the next Generation of Recommender Systems: A Survey of the State-of-the-art and Possible Extensions." IEEE Trans. Knowl. Data Eng. IEEE Transactions on Knowledge and Data Engineering: 734-49. Print.

[3] Sauter, Vicki Lynn, and Vicki Lynn Sauter. Decision Support Systems for Business Intelligence. 2nd ed. Hoboken, N.J.: Wiley, 2010. Print.

[4] R.Bell, Y. Koren and C. Volinsky. The BellKor 2008 Solution to the Netflix Prize. 2008

[5] Said, Alan, Brijnesh J. Jain, Sascha Narr, and Till Plumbaum. "Users and Noise: The Magic Barrier of Recommender Systems." User Modeling, Adaptation, and Personalization Lecture Notes in Computer Science: 237-48. Print

[6] Plan a ride with Surface, Directions, and Turf.js (Mapbox) https://www.mapbox.com/blog/dc-bikeshare-revisited/