"Sistema de predicción y recomendación personalizada basada en ranqueo de ítems homogéneos usando filtrado colaborativo"

Fausto Ruiz[1], Alejandro Díaz[2], Hugo Chang[3] , Fabricio Echeverría[4]

[1] Ingeniero en Computación Sistemas de Información 2006
[2] Ingeniero en Computación Sistemas de Información 2006
[3] Ingeniero en Computación Sistemas Tecnológicos 2006
[4] Director de Tópico, Magister en Sistemas de Información Gerencial, Escuela Superior Politécnica del Litoral, 2006, Postgrado Ecuador, ESPOL, 2006, Profesor desde 2000.

Resumen

Los sistemas de recomendación utilizan técnicas de descubrimiento de conocimiento al problema de recomendaciones personalizadas de información, productos o servicios durante una interacción en tiempo real.

Estos sistemas, especialmente los basados en el filtrado colaborativo del k-vecino más cercano están alcanzando un amplio éxito en la Web. El tremendo crecimiento en la cantidad de información disponible y el número de visitantes a los sitios Web en los recientes años propone algunos retos claves para estos sistemas. Estos retos son: producir recomendaciones de alta calidad, generar muchas recomendaciones por segundo para millones de usuarios y artículos, además de alcanzar una alta cobertura ante la dispersión de los datos.

En los sistemas tradicionales de filtrado colaborativo la cantidad de trabajo aumenta con el número de participantes en el sistema, por lo que son necesarias nuevas tecnologías en sistemas de recomendación que puedan producir rápidamente recomendaciones de alta calidad a gran escala.

Para lograr estas expectativas hemos explorado con técnicas de filtrado colaborativo basadas en ítems. Las técnicas basadas en ítems analizan primero la matriz ítem-usuario para identificar relaciones entre artículos diferentes, y luego usan estas relaciones para computar indirectamente recomendaciones para los usuarios.

En este artículo analizaremos cómo fue implementado nuestro sistema AFI Restaurantes y la optimización de los algoritmos utilizados para la predicción y generación de recomendaciones basadas en ítems.

Palabras Claves: filtrado colaborativo, k-medias, clusterización, sistema de recomendación, minería de datos.

Abstract

A recommendation system uses knowledge discovery techniques for solving the problem of making personalized recommendations of information, products or services during a real-time interaction.

These systems, specially the ones based on collaborative filtering of k-nearest neighbor, are reaching a great success on the Web. The huge growing in the amount of available information and the number of web site's visitors in recent years, propose critical challenges to the systems. These challenges are: produce high quality recommendations, generate lots of recommendations per second for millions of users and items and accomplishing a high support for data spreading.

In traditional collaborative filtering systems, the amount of work increase with the number of users participating in the system; that's why new technologies on recommendation systems are needed, so they can quickly provide high quality recommendations at a great scale level.

To accomplish these goals, we have explored the item-based collaborative filtering techniques. The item-based techniques first analyze the item-user matrix to identify relationships between different items, and then, use these relationships to compute, indirectly, recommendations for users.

In this article, we'll analyze how AFI Restaurantes, our recommendation system, was implemented and the algorithm optimization used to predict and generate item-based recommendations.

Key words: collaborative filtering, k-means, clustering, recommendation system, data mining.

1. Introducción

Actualmente, vivimos en un medio donde la oferta y demanda de productos y servicios se mueve a grandes velocidades. Este panorama representa un reto para las personas y/o empresas que ofertan sus productos, al tener que satisfacer las necesidades, cada vez más personalizadas, de los consumidores.

El consumidor se ve abrumado por el abanico de posibilidades que se le presenta al momento de elegir un determinado objeto cultural, entiéndase por ello a objetos de uso cotidiano para el esparcimiento, alimentación, entretenimiento, entre otros. Un problema evidente surge a la vista: cada vez será más difícil escoger un objeto cultural como una película, un libro o un restaurante, que sea realmente de nuestro interés y que, además de ir acorde a nuestro gusto, satisfaga nuestra expectativa de calidad. Usando una analogía, diríamos que es como "encontrar una aguja en un pajar".

¿Puede la tecnología ayudar de alguna manera a las personas a elegir realmente lo que buscan? ¿Ayudar a "encontrar la aguja en el pajar"? Pensamos que sí, y esto ha marcado el origen del presente proyecto de tesis.

Los sistemas de predicción y recomendación forman parte de un grupo de nuevas tecnologías que se encuentran dentro del estudio de la minería de datos, la cual intenta aprovechar el conocimiento que poseen ciertas personas acerca de determinados objetos culturales para ayudarse entre sí, generar nuevo conocimiento y darles a conocer nuevos objetos que posiblemente sean de su interés.

Los sistemas de predicción, hacen uso de mecanismos naturales de búsqueda que seguro hemos utilizado; en más de una ocasión, hemos tomado decisiones sobre la elección de un producto o servicio basado en la opinión de una persona o grupo de personas en quienes confiamos, de manera que objetos recomendados por otros (un libro, una película, una obra musical) nos estimulan a conocerlos.

1.1. Objetivos del proyecto

- Exponer las características y funcionamiento de los algoritmos de predicción basados en ítems y, tomando como fundamento las técnicas de la minería de datos, poder mostrar su aplicación específica en la tarea de predicción y recomendación.
- Desarrollar e implementar un sitio Web que se base en el uso de algoritmos de filtrado colaborativo y solucione el problema de hacer recomendaciones personalizadas de ítems de un mismo tipo (homogéneos) a usuarios registrados que presentan comportamientos y gustos similares. El dominio de producto o servicio escogido es el de los restaurantes de Guayaquil.
- Definir las ventajas y retos del uso del filtrado colaborativo en la elaboración de recomendaciones, así como también métodos y sugerencias para garantizar la escalabilidad en ambientes de producción que manejan grandes volúmenes de datos.

2. Sistemas de recomendación basado en filtrado colaborativo

Un sistema de recomendación basado en un filtro colaborativo podría definirse como: "aquel sistema en el que las recomendaciones se realizan basándose solamente en los términos de similitud entre los usuarios". [1]

Para cada usuario se crea un conjunto de "vecinos cercanos", usuarios cuyas evaluaciones anteriores tienen grandes semejanzas a las del usuario en cuestión. Los resultados para los elementos no calificados se predicen en base a la combinación de puntos (scores) conocidos de los vecinos cercanos.

En el filtrado colaborativo, el sistema no analiza los elementos evaluados, sino que las recomendaciones se basan solamente en la similitud entre usuarios.

2.1. Representación de un sistema de recomendación

En un sistema de recomendación colaborativo habrá que representar por una parte los objetos del sistema, y por otra parte tendremos que representar de alguna forma a los usuarios. De esta manera el sistema recomendará al usuario en cuestión, objetos del sistema que no conozca y a los que otros usuarios muy parecidos a él han valorado positivamente. [2]

Consideremos en primer lugar un conjunto O, formado por todos los objetos del sistema:

$O = \{O_1, O_2, O_3, \ldots, O_i, \ldots, O_m\}$

Por otra parte, tendremos otro conjunto U constituido por todos los usuarios:

$U = \{U_1, U_2, U_3, \ldots, U_i, \ldots, U_n\}$

Siguiendo esta representación, cada usuario U_i del sistema podríamos entenderlo como un vector

$U_i = (Calif_{i1}, Calif_{i2}, Calif_{i3}, \ldots, Calif_{ij}, \ldots, Calif_{in})$

donde cada componente $Calif_{ij}$ representaría la calificación con la que el usuario U_i ha valorado al objeto O_j.

Un usuario del sistema no tiene que haber calificado a todos los objetos existentes; es más, si esto fuera así, no tendría sentido el sistema de recomendación puesto que no tendría qué recomendar al usuario puesto que ya lo conoce todo.

Por lo tanto, dentro de este vector U_i que representa al usuario, habrá componentes vacías, componentes que el usuario no ha votado debido a que aún no las conoce. Tenemos entonces un espacio vectorial un poco peculiar, con vectores que no tienen valor en todas sus componentes. Este aspecto influirá en la manera en que tendremos que definir las diferentes funciones entre vectores que utilizaremos.

2.2. Representación de nuestro sistema de recomendación

La información que representa a los objetos y a los usuarios de nuestro sistema de recomendación reside en la base de datos de AFI Restaurantes.

Los objetos que componen nuestro sistema existen como registros de una tabla de la base de datos que denominamos Restaurante. Por otra parte, los datos correspondientes a los usuarios (nombre, etc.) constituyen los registros de otra tabla de la base de datos, la tabla Usuario. Como señalamos anteriormente, un usuario del sistema debe representarse como un vector formado por las votaciones de los restaurantes que haya votado, teniendo vacías estos vectores aquellas componentes que representen a los objetos que no haya votado. (Ver Figura 1 - Ejemplo de representación)

Tabla Usuario:

Usuario
Cristina
Xavier

Tabla Restaurante:

Restaurante
Mc Donald's
Friday's
KFC

Tabla Calificacion:

Usuario	Restaurante	Calificacion
Xavier	Mc Donald's	4
Cristina	Friday's	4,5

Figura 1 - Ejemplo de representación

Para representar estos vectores de usuarios (Ver Figura 1 - Ejemplo de representación), existe una tabla más en la base de datos del sistema, denominada Calificación que estable una relación directa Usuario-Restaurante, en la que cada registro representará una votación de un objeto concreto por un individuo.

Para realizar las recomendaciones a los usuarios en estos sistemas colaborativos representados con un espacio vectorial de este tipo, el sistema elegirá objetos que el usuario desconozca y que tengan una valoración alta en otros usuarios parecidos.

Debemos definir entonces métodos que nos sirvan para conocer el grado de similitud entre dos usuarios, ó lo que es lo mismo, grado de similitud entre sus dos vectores representativos. Medidas que pueden servir para este objetivo es el coseno del ángulo formado por ambos vectores.

Debido a que nuestro espacio vectorial es algo peculiar (puesto que tenemos componentes en los vectores sin valor en algunos de ellos) sólo consideraremos para cualquier cálculo las componentes de los vectores para las que existe valor en ambos.

2.3. Ejemplo de representación de AFI Restaurantes

A continuación pondremos un ejemplo, relacionado a nuestro contexto de restaurantes, que aclare los conceptos explicados hasta ahora.

Supongamos que tenemos un sistema de recomendación colaborativo donde el conjunto de objetos a recomendar son los restaurantes de la ciudad de Guayaquil.

Suponiendo que:

- El Usuario1 (U_1) ha frecuentado 6 restaurantes: A, B, D, J, L, V; y ha calificando respectivamente dichos restaurantes con: 9.5, 3, 3, 8, 7 y 4.

- El Usuario2 (U_2) ha frecuentado por su parte 3 restaurantes: C, J, V; a los que en este caso dio la siguiente valoración: 10, 5 y 6.

Para representar a estos dos usuarios utilizaríamos los siguientes vectores de usuario:

U_1 = (9.5, 3, vacio, 3, vacio, …, vacio, 8, vacio, 7, vacio, ..., vacio, 4, vacio, ..., vacio)

U_2 = (vacio, vacio, 10, vacio, ..., vacio, 5, vacio, ..., vacio, 6, vacio, ..., vacio)

De esta manera, vemos que quedarán vacías las coordenadas de un vector que correspondan a libros que desconoce el usuario en cuestión.

A manera de ejemplo adicional, diríamos que para la representación de la Figura 1, los vectores teóricos correspondientes a los usuarios Xavier y Cristina serían:

Xavier = (4, vacio, vacio)

Cristina= (vacio, 4.5, vacio)

3. AFI Restaurantes y minería de datos

3.1. *Técnicas de minería de datos escogidas*

Con el fin de resolver el problema planteado, dado el modelo de extracción de conocimiento que utilizamos, se usan dos tareas de minería de datos:

1. **Clasificación**: Esta tarea se utiliza con el propósito de obtener temporalmente una lista ordenada, de los usuarios más "similares" entre sí. Tarea que se va a solucionar usando el método de *vecinos más próximos o KNN*. [3]
2. **Agrupamiento:** Esta tarea se utiliza para generar y almacenar, persistentemente, grupos de usuarios con características similares con el fin de abreviar la carga computacional, reduciendo de esta manera la cantidad de información procesada. El método que escogimos para determinar estos grupos es el de K-medias. [3]

Además de las técnicas específicas de minería de datos utilizadas, fue crucial elaborar un programa que nos permita gestionar el desarrollo de "AFI Restaurantes" como un proyecto de minería de datos. Para dicha formulación, usamos el **modelo y guía de referencia CRISP-DM [4],** estándar de la industria para el proceso de minería de datos, que nos permite definir: qué es necesario incluir en el programa, qué estructura se va a seguir y cuál será su extensión y alcances.

3.2. *Algoritmos de análisis multivariante escogidos*

En la implementación de "AFI Restaurantes", nuestro sistema de predicción y recomendación se han introducido los siguientes algoritmos:

1. **Algoritmo KNN (vecino más cercano):**

Permite el descubrimiento de conocimiento no asistido y pertenece a los métodos sin modelo, y retardados o perezoso (*lazy*) que se basa en actuar para cada pregunta o predicción requerida.

Este algoritmo consigue una lista ordenada de los usuarios más similares al usuario al que se desea entregar una recomendación. Implica una tarea de comparación de un usuario contra todos, el cual representa un procedimiento de alta demanda computacional y usaremos las técnicas que se mencionan a continuación para complementar el trabajo de recomendación. Un ejemplo de aplicación se verá más adelante en la sección 3.4 ***Ejemplo de aplicación de los algoritmos escogidos***.

2. **Algoritmo de K-medias:**

Este tipo de método es conveniente utilizarlo cuando los datos a clasificar son muchos y/o para refinar una clasificación obtenida utilizando un método jerárquico. Supone que el número de grupos es conocido a priori.

Existen varias formas de implementarlo pero todas ellas siguen, básicamente, los siguientes pasos:

1) Se seleccionan k centroides o semillas donde k es el número de grupos deseado.
2) Se asigna cada observación al grupo cuya semilla es la más cercana.
3) Se calculan los puntos semillas o centroides de cada grupo.
4) Se iteran los pasos 2) y 3) hasta que se satisfaga un criterio de parada como, por ejemplo, los puntos semillas apenas cambian o los grupos obtenidos en dos iteraciones consecutivas son los mismos.

Este algoritmo genera grupos de usuarios con características similares, dichos grupos se mantendrán almacenados en la base de datos para posterior uso y además formarán parte de un histórico de grupos en nuestro modelo multidimensional de base de datos.

3.3.Implementación de los algoritmos

La finalidad de usar el algoritmo de K-medias es reducir la carga computacional al momento de dar recomendaciones y de resolver posibles problemas de escalabilidad.
Para dicho efecto, el algoritmo K-medias hará uso de los siguientes procedimientos que mejoran su desempeño:

El test F de reducción de variabilidad
Al hacer uso del algoritmo de K-medias es necesario fijar el número de grupos que se van a generar. Este número no se puede estimar con un criterio de homogeneidad porque de ser así se formaría un solo grupo; por lo tanto, usamos el test F que estima el número de grupos óptimo que se necesita crear. El test F le da un carácter dinámico a la generación de grupos de acuerdo a la variabilidad de los datos.

Algoritmo de Fisher
Realiza la tarea de predicción. Este procedimiento permite asignar un usuario a un determinado grupo que ha sido creado con anterioridad. Al predecir a qué grupo un usuario será asignado logramos que la tarea de recomendaciones mejore considerablemente en su tiempo de respuesta.
En el método de Fisher, la obtención de las distintas funciones se deriva de un proceso de obtención de raíces y vectores propios de una forma cuadrática. La suma de cuadrados entre grupos de cada función discriminante, viene definida por un autovalor $\lambda(i)$.
El algoritmo KNN (recomendación) se ejecutará comparando el usuario contra su grupo asignado y sus centroides; y ya no contra todos los usuarios registrados en el sistema.

3.4.Ejemplo de aplicación de los algoritmos escogidos

En un principio, consideraremos un número pequeño de usuarios que usan "AFI Restaurantes". En este escenario, el Usuario A pide una recomendación y el sistema hace uso del algoritmo de vecino más próximo para generarla. Debemos tener en cuenta que, para al momento, no existen grupos de usuarios definidos, de manera que el proceso involucra todos los usuarios existentes.

Conforme se incrementa el número de usuarios, el tiempo necesario para generar las recomendaciones para el Usuario A aumentará. Para solucionar este problema de escalabilidad, el Administrador de "AFI Restaurantes" ejecuta periódicamente el algoritmo k-medias, que en su procedimiento usa el test F, para construir la cantidad óptima de grupos entre los usuarios existentes. Los grupos son almacenados en la base de datos.

La próxima vez que el Usuario A solicite una recomendación, se usa el algoritmo de Fisher para predecir a qué grupo pertenece, y se prosigue a usar el algoritmo de vecino más cercano sobre los usuarios de dicho grupo, ya no involucrando a todos los usuarios existentes.

4. Conclusiones y recomendaciones

- El proyecto de "AFI Restaurantes" empezó como una idea planteada por uno de lo integrantes del grupo quien había escucha de sistemas de predicción y recomendaciones durante los cursos de Ingeniería de Software dictados en la FIEC.

- El tema cobró verdadero interés durante el tópico de graduación de "Aplicaciones de minería de datos" dirigido por el Msc. Fabricio Echeverría, donde se empezó a trabajar la idea para ser presentada como tesis de grado.

- El proyecto empezó siendo algo pequeño, sin embargo, nuevas ideas y profundidad en los temas se fueron mostrando conforme avanzó el desarrollo del tópico. El desarrollo de las etapas de análisis y diseño del sistema fue guiado por el director del tópico y revisadas, cuidadosamente, por los integrantes del grupo; presentándose continuamente, hasta el final, cambios y sugerencias en el contenido y diagramación.

- Un punto importante a considerar fue la necesidad de investigar soluciones para resolver los problemas de escalabilidad a los que está sujeto este tipo de sistemas; aspecto que no era de nuestro conocimiento al momento de embarcarnos en el desarrollo de este tema de tesis.

- Durante la implementación de nuestro sistema "AFI Restaurantes" pudimos notar que existe un amplio espectro de aplicaciones vinculado a la minería de datos y, particularmente relacionadas al análisis de conglomerados o clústeres. Con el uso correcto de estas herramientas, hemos descubierto un potencial con el que antes solo podíamos imaginar, la fusión del poder de procesamiento de estos días en conjunto, con el poder de análisis matemático de la Estadística, permiten obtener actualmente información valiosa e interesante para pre y post análisis; lo que antes sólo era un historial o repositorio de datos, se vuelve ahora información de relevancia que ayuda a tomar decisiones.

- Recomendamos siempre centrarse en un buen diseño o perspectiva al atacar algún problema en particular, siempre remitiéndose a la teoría o mejores prácticas para obtener resultados esperados o acordes en tiempos esperados o adecuados. El modelo y guía de referencia CRISP-CM es recomendado como estándar para la planeación, organización, dirección y control de un proyecto de minería de datos.

- Recomendamos que la herramientas de desarrollo y algoritmos que se utilicen, tengan optimizadas las operaciones de minería de datos; en la misma línea recomendamos el uso de paralelismo o hilos (*threads*) en los procesos del análisis de conglomerados, proporcionalmente al tamaño de su repositorio de datos, para mantener el rendimiento y desempeño de sus aplicaciones desarrolladas.

5. Agradecimientos

Agradecimientos a nuestro director de tesis, Msc. Fabricio Echeverría, por su constante estimulación para que todos estos conceptos estudiados tuvieran efectiva aplicación en el mundo real que, paradójicamente, son de ahí de donde vienen. A sí mismo, a todas las personas que colaboraron con información específica y a nuestros compañeros que han hecho de este ciclo una experiencia enriquecedora y amena para nuestra vida profesional.

6. Referencias

[1] M. Balabanovic and Y. Shoham, (1997) “Content-Based Collaborative Recommendation,” Comm. ACM, Mar.1997, pp. 66-72.
[2] Premios NAI. Memoria SRI, Sistema Inteligente de Recomendaciones (pdf). 2003.
[3] Hernández, José. Introducción a la Minería de Datos. Pearson. 2004
[4] CRISP-DM. Cross Industry Standard Process for Data Mining. [en línea], disponible en: http://www.crisp-dm.org/Process/index.htm