Análisis eficiente de Transacciones para la mejora de Recomendaciones Web

E. Lazcorreta F. Botella A. Fernández-Caballero

XIII Congreso Internacional de Interacción Persona Ordenador
(Interacción'2012)
3-5 octubre 2012, Elche, España

CENTRO DE INVESTIGACION OPERATIVA
I3A

1. Introducción
2. Tipos de transacciones
3. Análisis eficiente de transacciones tipo II
   - Refinamiento
   - Ítems no relacionados
4. Experimentación
5. Conclusiones y trabajo futuro

CENTRO DE INVESTIGACION OPERATIVA
I3A

Si trabajamos con grandes colecciones de datos y queremos obtener en poco tiempo la información más relevante que contienen podemos recurrir a la extracción de patrones mediante Minería de Datos.

Entre los patrones más utilizados están las Reglas de Asociación, que se basan en la co-ocurrencia de ítems en grandes colecciones de transacciones.

Tras muchos ensayos hemos encontrado un tipo de transacciones que pueden ser tratadas de un modo más eficiente que el que se han venido utilizando hasta hoy día. Aplicando una transformación a las transacciones en estudio podemos usar cualquiera de los algoritmos existentes de Minería de Reglas de Asociación para obtener más información en menos tiempo y utilizando menos recursos.

En nuestros primeros ensayos hemos obtenido tiempos de ejecución muy rápidos, lo que nos permitirá mejorar los tiempos de respuesta de un Sistema de Recomendación Web.

CENTRO DE INVESTIGACION OPERATIVA
I3A

Una transacción no es más que un conjunto de elementos que están agrupados por algún motivo sobre el que queremos investigar.

La transacción más clásica es la denominada "cesta de la compra", que recoge los artículos adquiridos por un mismo usuario en una compra o bien en todo su historial de compras, según el tipo de negocio que estemos analizando (un supermercado o una librería, p.ej.). ARM pretende obtener de forma rápida la mayor cantidad de co-ocurrencias observables en un gran almacén de repositorios, lo que puede aportar información de valor a los gestores del negocio.

El auge y la eficiencia obtenida con la ARM ha extendido el concepto de transacción mediante el modelado de otro tipo de datos que inicialmente no son transacciones. Es el caso de las mediciones hechas sobre un número determinado (y codificado) de características medibles en una gran cantidad de individuos, p.ej. las características de color, textura. . . de un gran número de setas. Estas "transacciones" se analizan mediante métodos rápidos de Asociación y el resultado obtenido se aplica en disciplinas como la Clasificación.

Tras varios años de experimentación con grandes colecciones públicas de datos hemos descubierto que si trabajamos con este tipo de transacción no obtenemos parte de la información que contienen porque no usamos la información que ya tenemos sobre esas transacciones: si una seta es marrón no puede ser roja, existe una información estructural en este tipo de transacciones que la diferencia de las transacciones clásicas.

En este trabajo proponemos incluir esa información estructural en el propio análisis de ARM para eliminar la información redundante que ya podemos conocer antes de llevar a cabo cualquier análisis.

En nuestros primeros experimentos hemos descubierto que no sólo se reduce notablemente el tiempo y recursos necesarios para el análisis de este tipo de transacciones sino que también podemos obtener información que con los métodos clásicos no se obtenía. Además podemos utilizar cualquier algoritmo de ARM, aprovechando cada una de sus ventajas.

Provisionalmente hemos llamado a este tipo de modelización de datos "transacciones de tipo II".

Lo que nos hizo pensar en la existencia de más de un tipo de transacciones fue el hecho de almacenes de datos públicos y usados en muchos artículos de investigación sobre ARM no eran excesivamente grandes pero presentaban serias dificultades para un análisis completo. El *dilema del ítem raro* aparecía con frecuencia en estas colecciones "no demasiado grandes" de datos.

Al observar con más detalle los resultados obtenidos encontramos que en estos almacenes de datos existía un alto índice de co-ocurrencia entre todos sus elementos. Y esta característica es la que provoca que ARM no sea altamente eficiente para tratar a este tipo de datos.

Si ya sabemos que la cualidad `color = marrón` no puede estar en una transacción que contiene la cualidad `color = rojo` ¿Por qué hemos de utilizar tiempo y recursos en comprobarlo a través de los datos que tenemos?

Al tratar transacciones de este tipo podemos prescindir del estudio de las setas rojas, p.ej., pues ya sabemos que si una seta no es de ninguno del resto de colores que hemos codificado será necesariamente roja. De este modo eliminamos del almacén de datos todos aquellos elementos que pueden ser deducidos a partir de la existencia de otros elementos que

Podemos reducir al máximo el número de elementos del almacén de transacciones si eliminamos los elementos que aparecen con mayor frecuencia, con lo que el tiempo de ejecución y los recursos empleados en el análisis se reducen al máximo sin perder información.

También podemos seguir otros criterios para decidir qué valor de un atributo debe desaparecer de la colección inicial de datos, p.ej. la "simpatía" de un elemento (el número de elementos distintos con el que se relaciona).

El análisis de nuestros primeros resultados puede aportarnos mayor información para conseguir un refinamiento más adaptado a cada colección de datos.

Inesperadamente nos hemos encontrado con un tipo de información que los algoritmos clásicos por sí solos no son capaces de extraer.

La reducción de recursos necesarios para el análisis conlleva que en algunos casos podamos estudiar el 100 % de las co-ocurrencias existentes en un almacén concreto de datos. Con el análisis clásico esto no es posible y obliga a estudiar sólo los elementos que tienen un soporte mínimo, despreciando la información de los elementos que aparecen con menor frecuencia.

Sin embargo al extraer TODA la información de Asociación que contiene el almacén de datos podemos asegurar que hay elementos que no se relacionan con otros.

En el artículo se muestran tablas con comparativas entre los resultados obtenidos usando métodos clásicos de ARM y los obtenidos considerando que se está trabajando con transacciones de tipo II.

Dichas comparativas muestran que esta pequeña consideración es suficiente para obtener mayor información en mucho menos tiempo, pudiendo operar con los resultados prácticamente en tiempo real.

Estimamos que su aplicación a sistemas que requieran de análisis en tiempo real, como los Sistemas de Recomendación Web, puede aportar grandes mejoras en el servicio.

Incorporar a un análisis “ciego” como es el ARM la información que ya poseemos sobre los datos con los que estamos trabajando mejora notablemente el rendimiento de los algoritmos ya conocidos.

Actualmente estamos finalizando un software que detecta el tipo de transacción que se está analizando y decide por sí mismo si se ha de hacer una reducción de los datos a analizar en los primeros pasos del algoritmo.

Gracias por vuestra atención.

¿...?