Recomendaciones en sistemas web mediante el estudio de ítems raros en transacciones

Enrique Lazcorreta
Inst. Univ. Centro de Investigación Operativa (CIO)
Universidad Miguel Hernández
03202 Elche
enrique@umh.es

Federico Botella
Inst. Univ. Centro de Investigación Operativa (CIO)
Universidad Miguel Hernández
03202 Elche
federico@umh.es

Antonio Fernández-Caballero
Inst. Investigación en Informática de Albacete (I3A)
Universidad de Castilla-La Mancha
020071 Albacete
caballer@dsi.uclm.es

Resumen

Las recomendaciones en portales web se enriquecen cuando incorporan información sobre el uso real del portal. Una de las herramientas que proporcionan dicha información es el minado de reglas de asociación (ARM) en las sesiones de navegación de los usuarios a través del portal. Si el número de páginas del portal es muy grande la ARM se tropieza con el dilema del ítem raro, que postula que no se puede encontrar información sobre las páginas poco visitadas sin obtener una explosión de reglas de asociación. Este dilema ha sido estudiado desde una perspectiva que sobrecarga las tareas del algoritmo usado y de sus analistas y no aporta reglas de uso frecuente. En este artículo se introduce un nuevo enfoque al dilema que permite a los algoritmos de ARM encontrar simultáneamente información de interés sobre los páginas del portal poco visitadas, con un consumo mínimo de recursos, sin intervención de los analistas y aportando conocimiento útil.

1. Introducción

Desde la irrupción de la informática en pequeñas y grandes empresas se puede guardar en bases de datos todo tipo de información, entre otras las secuencias de navegación de un usuario a través de un portal web. El volumen de datos almacenados sobre las páginas solicitadas al portal crece a diario por lo que su análisis inicial solo puede hacerse usando técnicas de Minería de Datos (en adelante, MD) que permiten extraer el conocimiento que encierran grandes colecciones de datos.

Las transacciones son agrupaciones de ítems hechas por los usuarios de un servicio, por ejemplo las distintas compras realizadas por un usuario en un comercio electrónico, las paginas solicitadas por un usuario en una visita a un portal web o los recursos utilizados por un alumno en un portal de e-learning, siempre que no tengamos en cuenta el orden en que han ocurrido. Llamando itemset a un conjunto cualquiera de ítems (entendiendo por ítem cada una de las páginas del portal web), si analizamos las transacciones realizadas por los usuarios del portal podemos observar la existencia de itemsets que se repiten, itemsets que llamaremos patrones ya que pueden ser usados para comprender el comportamiento de los usuarios del sistema. En tal caso podemos pensar que existe algún tipo de relación entre los ítems que forman dichos itemsets, relación que llamaremos asociación de ítems y que, recordemos, se producen por el uso del servicio y no por su diseño.

Cuando se trabaja con pocos ítems o con pocas transacciones el análisis que procede es el estudio de la correspondencia entre los ítems que componen las transacciones. Sin embargo son muchos los servicios que proporcionan gran cantidad de ítems y recogen a diario miles de transacciones cuyo estudio desde la perspectiva del análisis estadístico de la correspondencia es inabordable. Para resolver este problema la MD proporciona técnicas tratables computacionalmente, como el minado de reglas de asociación (ARM, por sus siglas en inglés *Association Rule Mining*).

Las reglas de asociación son expresiones del tipo $X \rightarrow Y$, donde X e Y son itemsets que no poseen ítems en común. La calidad de las reglas se mide por su soporte (porcentaje de veces que se verifica la regla en las transacciones) y su confianza (porcentaje de veces que una transacción que contiene el itemset X también contiene el itemset Y). La irrupción de la MD en el estudio de las reglas de asociación se debe al altísimo número de reglas que pueden ser generadas y la imposibilidad de ser tratadas computacionalmente con la tecnología actual y un enfoque puramente estadístico.

Sin embargo, los almacenes de transacciones (*D*

en adelante) muy grandes proporcionan tal cantidad de reglas que su tratamiento se vuelve inabordable incluso con la MD. Para afrontar este problema hay que recurrir al concepto del soporte mínimo en el ARM sobre grandes bases de datos, [1], categoría de la MD abundantemente estudiada hasta la fecha en [2], [3], [4] y [5]. Entre otras aplicaciones, la ARM es fuente de información para los sistemas de recomendación de páginas web, artículos de comercio y contenido educativo, propuestos en [6], [7] y [8].

El principal inconveniente a la hora de abordar una solución al problema de ARM es el crecimiento exponencial de resultados a estudiar ya que un conjunto I de m ítems distintos puede producir hasta 2^m itemsets diferentes. Aunque no todos los itemsets estén presentes en D, para poder contar su soporte hemos de generar el conjunto completo de soluciones posibles, donde poder anotar el número de veces que aparece cada uno de ellos. La definición de soporte mínimo (minSup) permite aliviar este inconveniente pues supone una reducción del número de itemsets a estudiar: sólo aquellos cuya frecuencia sea superior a un umbral fijado por el analista. La mejor aportación a este problema fue propuesta en [9] mediante el algoritmo Apriori, que genera únicamente los candidatos a itemset frecuente que pueden alcanzar el soporte mínimo por no contener ningún itemset infrecuente.

El problema llamado dilema del ítem raro [10] surge cuando queremos obtener reglas de asociación para todos los ítems de I que se relacionen mayoritariamente con algún otro ítem en D. Si bajamos el umbral del soporte mínimo para obtener información sobre los ítems poco frecuentes en D es muy probable que desbordemos la capacidad del ordenador en que se ejecute el algoritmo, y en caso de ser ejecutado por completo obtendremos una enorme cantidad de reglas que no son realmente importantes, por ejemplo aquellas que relacionan con poco soporte dos ítems muy frecuentes en D.

En este trabajo mostramos como sin un coste computacional excesivo ni mayor dedicación de los analistas se puede modificar el algoritmo Apriori para que encuentre reglas interesantes sobre ítems que no son frecuentes sin provocar una explosión de reglas. Las reglas encontradas se utilizarán en mayor medida si cambiamos la perspectiva desde la que se observan.

En la sección 2 se expone el problema, las soluciones propuestas hasta la fecha y una propuesta para su solución, en la sección 3 se muestra una aplicación a las recomendaciones web y en la sección 4 se presentan las conclusiones de este trabajo y el trabajo futuro.

2. Estudio de ítems raros

Los ítems raros son ítems que por alguna razón no aparecen con frecuencia en las transacciones. En una cesta de la compra los ítems raros suelen ser muy caros o duraderos, y también pueden ser novedades que tardarán en ser comprados de forma frecuente. En un portal web los ítems raros son páginas muy especializadas, de baja calidad o páginas de novedades.

Los ítems exclusivos (por su alto precio, larga durabilidad o especialización) serán siempre ítems raros aunque se incorporen en un sistema de recomendación, ya que son infrecuentes por naturaleza. Sin embargo, los ítems nuevos deberían ser incorporados a los sistemas de recomendación hasta que alcancen la categoría de ítems frecuentes, momento en el cual ya pueden ser tratados mediante la búsqueda clásica de reglas de asociación.

Para abordar el problema del ítem raro [10] propone usar múltiples soportes, de modo que la relación entre dos ítems frecuentes sea considerada sólo si es una relación muy frecuente en D. De este modo se alivia considerablemente la carga de memoria requerida por el algoritmo y permite abordar el estudio sobre un número mayor de ítems. La asignación de soporte a cada ítem puede hacerse por parte del analista o bien teniendo en cuenta el propio soporte de cada ítem en D. En [11] se propone una modificación interesante del algoritmo propuesto en [10], incorporando al estudio medidas de tendencia de los ítems de D.

Con la primera propuesta de [10] se pueden incorporar los ítems nuevos asignándoles un soporte mínimo muy bajo mientras sean nuevos, sin embargo el analista debe decidir en algún momento cuándo debe modificar su soporte mínimo y qué nuevo soporte asignarle.

En las restantes propuestas el soporte mínimo se obtiene en función del soporte real del ítem, con lo que muchos ítems raros se incorporan al sistema con facilidad, sin embargo siguen quedando muchos ítems sobre los que no obtenemos información pues si intentamos incorporarlos al estudio se detiene la ejecución del algoritmo por falta de recursos. Además las reglas que proporcionan

sobre ítems poco frecuentes tienen bajo soporte y, en consecuencia, baja utilidad.

En base a las propuestas de [10] y [11] hemos realizado experimentos sobre repositorios medianos y grandes. Con los repositorios T10I4D100K y T40I10D100K, que contienen 870 y 942 ítems diferentes respectivamente, se ha logrado información sobre todos sus ítems reduciendo la información a almacenar sobre relaciones infrecuentes de ítems frecuentes (con el método clásico no es posible obtener dicha información con la tecnología de 32 bits). Con repositorios grandes hemos comprobado que no es posible abordar el estudio de todos sus ítems desde la perspectiva de las reglas de asociación, como ocurre con el repositorio kosarak, que contiene 41.270 ítems distintos. El algoritmo clásico permite obtener reglas de asociación únicamente entre 1.544 de sus ítems mientras que usando soporte múltiple podemos obtener información sobre 4.636 ítems, una gran mejora pero que deja un 89% de ítems del repositorio sin ningún tipo de información. Este porcentaje sería peor si el número de ítems en estudio fuera mayor, lo que no es anormal en muchos de los portales web existentes.

Si queremos obtener información útil de todos los ítems de *D* hemos de prescindir inicialmente del concepto de soporte mínimo. Pero no podemos obviar la capacidad de los ordenadores en que se ejecutará el algoritmo por lo que hemos de optar por un algoritmo tipo FP-Growth. Si además queremos obtener reglas útiles tendremos que cambiar la perspectiva desde la que se miran las reglas de asociación, lo que nos conduce a definir las *Reglas de Oportunidad* (RO).

Las RO son reglas cuyo antecedente es frecuente por lo que serán usadas con frecuencia en un sistema de recomendación, y cuyo consecuente es infrecuente por lo que permitirán recomendar ítems poco frecuentes a los usuarios.

El algoritmo 1 muestra cómo leer *D* sin generar candidatos (del mismo modo que se trabaja con la estructura FP-Tree), debido a que los grandes repositorios de transacciones pueden contener tantos ítems diferentes que el enfoque clásico de Apriori impediría la simple generación de C_2 (la colección de datos kosarak tiene 41.270 ítems distintos, que generan 851.627.085 candidatos a 2-itemset). La función `Incrementa` comprueba si existe el ítem y lo crea en caso necesario, de este modo se emplea mayor tiempo en la obtención de los 2-itemsets pero se evita una explosión de candidatos a considerar que haría abortar la ejecución del algoritmo por falta de recursos.

```
Input: D, sm (soporte mínimo) y om
       (oportunidad mínima)
Output: RO (Reglas de Oportunidad)
      y RA (Reglas de Asociación)
      presentes en D
/* Obtener frecuencia de todos los
ítems y 2-itemsets de D */
foreach transacción Ti en D
 foreach i1 en Ti {
  Incrementa(FP1[i1]);
  foreach (i1; i2) en Ti
   Incrementa(FP1[i1]->FP2[i2]);
 }
/* Extraemos las RO */
foreach i1 en FP1
 if (FP1[i1] >= sm) then
  foreach i2 en FP1[i1]->FP2
   if (FP1[i2] < sm y FP2[i2]/FP1[i1]
       >= om) then
    Añadir RO(it1 → it2);
```

Algoritmo 1: ORFind - Algoritmo de búsqueda de Reglas de Oportunidad

Al aplicar el algoritmo clásico tras la ejecución del algoritmo 1 a los repositorios T10I4D100K, T40I10D100K y kosarak hemos obtenido información de interés sobre todos los ítems de cada repositorio en un tiempo inferior al empleado con la implementación de las propuestas de [10] y [11]. Las reglas de asociación obtenidas entre ítems frecuentes tienen el suficiente soporte como para ser utilizados como patrones de comportamiento del colectivo estudiado. Las reglas de oportunidad obtenidas vinculan los ítems infrecuentes a los frecuentes, permitiendo a un sistema de recomendación saber cuál es el momento más oportuno para recomendar un ítem infrecuente de cuyo uso no tenemos aún información suficiente.

3. Recomendaciones web mediante el estudio de transacciones

En un Sistema de Recomendación Web se utiliza información sobre taxonomías, contenido semántico y uso del portal para hacer sugerencias a sus usuarios. Sin embargo cuando un ítem es nuevo no puede obtenerse información de su uso hasta que no logra un cierto soporte. Las reglas de oportunidad permiten incorporar en el sistema la información de uso de los ítems nuevos de modo

automático.
Las RO contienen los ítems frecuentes que son visitados conjuntamente con los ítems sin soporte mínimo. Esto nos permite sugerir un ítem raro a los usuarios que visitan alguna de las páginas que más favorecen la presencia de dicho ítem, p.ej. la página A.
Si la sugerencia es acertada es de esperar que el ítem raro vaya aumentando su soporte y se incorpore de modo natural al proceso de estudio de reglas de asociación, mostrando una asociación cada vez más fuerte a la página A.
Si la sugerencia no es acertada el ítem seguirá siendo raro y su asociación con la página A se irá diluyendo a favor de una asociación con otras páginas con las que realmente sea más afín. Esto se debe a que el número de veces que aparece el ítem raro en *D* es pequeño y cualquier aparición nueva del ítem modificará sustancialmente los porcentajes en que se basan los algoritmos de búsqueda de RA y RO.

4. Conclusiones

Las propuestas existentes para resolver el dilema del ítem raro reducen sustancialmente el número de ítems ignorados por el análisis de uso de un portal web. Sin embargo no proporcionan información de uso de un gran número de ítems. La información que proporcionan sobre un ítem de bajo soporte tiene escaso uso pues sólo se usará cuando el ítem sea requerido.
Las reglas de oportunidad permiten introducir de modo natural la información de uso de los ítems raros en los sistemas de recomendación. Esto da más visibilidad a los ítems raros y posibilita el incremento de su uso con lo que podría llegar a ser un ítem con el suficiente soporte como para entrar en el análisis tradicional (y más completo) de las reglas de asociación.
En la actualidad estamos comparando los resultados mostrados en este artículo con los que se obtienen al aplicar un híbrido entre la obtención de reglas de oportunidad y las propuestas realizadas en [10] y [11].

Agradecimientos

Este trabajo está financiado en parte por el proyecto nacional CICYT TIN2008-06596-C02-01

Referencias

[1] Agrawal, R., Imielinski, T., Swami, A. Mining association rules between sets of items in large databases. Proc. of the ACM SIGMOD International Conference on Management of data. Washington, D.C., United States, pp. 207-216, 1993
[2] Rozenberg , B., Gudes, E. Association rules mining in vertically partitioned databases. Data & Knowledge Engineering 59 (2), pp. 378-396. 2006
[3] Palshikar , G. K., Kale, M. S., Apte, M. M. Association rules mining using heavy itemsets. Data & Knowledge Engineering 61 (1), pp. 93-113. 2007
[4] Tseng , M.-C., Lin, W.-Y. Efficient mining of generalized association rules with non-uniform minimum support. Data & Knowledge Engineering 62 (1), pp. 41-64. 2007
[5] Lazcorreta, E., Botella, F., Fernández-Caballero, A. Towards personalized recommendation by two-step modied apriori data mining algorithm. Expert Systems with Applications 35 (3), pp. 1422-1429. 2008
[6] Kouris, I.N., Makris, C.H., Tsakalidis, A.K. Using information retrieval techniques for supporting data mining. Data & Knowledge Engineering 52 (3), pp. 353-383. 2005
[7] Botella, F., Lazcorreta, E., Fernández-Caballero, A., González, P., Mejora de la usabilidad y la adaptabilidad mediante técnicas de minería de uso web. Proc. of VI Congreso Interacción Persona-Ordenador. Thomson, 2005
[8] Han, J., Fu, Y. Discovery of multiple-level association rules from large databases. Proc. of Int. Conf. VLDB. pp. 420-431. 1995
[9] Agrawal, R., Srikant, R. Fast algorithms for mining association rules. Proc. of the 20th VLDB Conference, Morgan Kaufmann Publishers Inc., pp. 487-499. 1994
[10] Liu , B., Hsu, W., Ma, Y. Mining association rules with multiple minimum supports. Proc. of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, pp. 337-341. 1999
[11] Kiran , R. U., Reddy, P. K. An improved multiple minimum support based approach to mine rare association rules. IEEE Symposium on Computational Intelligence and Data Mining. 2009