Reglas de Oportunidad: mejorando las recomendaciones web

Enrique Lazcorreta[1], Federico Botella[1] y Antonio Fernández-Caballero[2]

[1] Instituto Universitario Centro de Investigación Operativa (CIO),
Universidad Miguel Hernándezde Elche
{enrique, federico}@umh.es
[1] Instituto de Investigación en Informática de Albacete (I3A),
Universidad de Castilla-La Mancha
caballer@dsi.uclm.es

Resumen. Los sistemas de recomendación web automáticos obtienen parte de información de los datos de uso del portal web. La búsqueda de reglas de asociación entre las páginas del portal a partir de las sesiones de sus usuarios es una de las fuentes más utilizadas para este propósito. El uso de soporte mínimo permite abordar el problema de la búsqueda de reglas de asociación pero impide obtener información sobre una gran cantidad de páginas del portal. En este artículo se introducen las *reglas de oportunidad*, que permiten a los algoritmos de búsqueda de reglas de asociación encontrar simultáneamente información sobre las páginas que no superen el soporte mínimo, con un consumo mínimo de recursos.

Keywords: Reglas de Asociación, Apriori, Sistemas de Recomendación, Minería de Uso de la Web.

1 Introducción

La búsqueda de reglas de asociación en grandes bases de datos, introducida en 1993 en [1] y abundantemente estudiada hasta la fecha [2, 3, 4, 5, 6], es fuente de información para los sistemas de recomendación de páginas web, artículos de comercio y contenido educativo, entre otros, propuestos en [2, 6, 7].

Con la búsqueda de reglas de asociación aplicada a los datos de uso de un portal web se pretende encontrar patrones de comportamiento que permitan mejorar la navegación a través del portal. Si un usuario solicita una página web se le puede sugerir que también visite otras páginas que los propios usuarios han visitado junto a la solicitada. De este modo son los propios usuarios quienes asocian las páginas que visita en una misma sesión, como si las estuvieran agrupando por tener algo en común.

Cuando son muchos los usuarios que agrupan las mismas páginas web podemos pensar que debe haber algún motivo para ello, pero cuando son pocos los que lo hacen podríamos llegar a pensar que la agrupación se ha hecho al azar, sin que exista una relación real entre las páginas agrupadas. De esta idea surge el concepto de soporte, el

número de veces que aparece repetido un grupo de páginas web. Si este número es pequeño, con respecto a cierto umbral fijado por el analista, no se generará una regla de asociación entre esas páginas según el enfoque clásico de la búsqueda de reglas de asociación.

En un sistema de recomendación de páginas web el uso del soporte penaliza a las páginas web que han sido visitadas pocas veces, entre otras las páginas web de reciente incorporación, pues no tienen suficiente soporte para formar parte de una regla de asociación que genere una sugerencia. Y conforme funciona el sistema de recomendación el problema puede seguir creciendo pues los enlaces sugeridos son siempre de páginas web que ya tienen suficiente soporte.

Los primeros trabajos en plantear el llamado *problema del ítem raro* proponían separar la base de datos en grupos de ítems con frecuencias homogéneas y estudiar cada grupo independientemente [8] o bien agrupar varios ítems infrecuentes en uno solo de modo que se incremente su frecuencia [9]. En el primer caso no se encuentran reglas que relacionen ítems dispuestos en diferentes grupos y en el segundo no se obtiene información referida a los ítems particulares que forman el ítem compuesto estudiado.

En [10,11,12] se propone utilizar múltiples umbrales. Si queremos facilitar la asociación de una página web con el resto podemos asignarle un umbral bajo. Sin embargo esta información adicional no se obtiene del uso real del sitio web. [10] presenta una modificación del algoritmo Apriori, MSApriori, en que cada ítem tiene su propio soporte mínimo, para obtenerlo primero comprueban el soporte observado en la base de datos y si un ítem no supera el soporte mínimo general se le asigna un soporte mínimo particular calculado como una fracción del soporte real del ítem. Una vez determinados los soportes mínimos de cada ítem ordenan los ítems en base a su soporte particular y después proceden con el algoritmo haciendo comparaciones con cada uno de los soportes involucrados en cada *k*-itemset. En [11] proponen el algoritmo CFP-growth basado en la estructura FP-tree. La mejora de la propuesta respecto a [10] está en la escalabilidad del algoritmo ya que en la primera propuesta un simple cambio de soporte mínimo aplicado a un ítem conlleva la re-lectura de toda la base de datos para obtener las nuevas reglas de asociación. En [12] se plantea el minado de reglas de asociación generalizadas bajo la influencia de una taxonomía.

Sin un coste computacional excesivo se puede modificar el algoritmo Apriori para que encuentre reglas interesantes sobre ítems que no son frecuentes.

En la sección 2 se expone el problema, en la sección 3 se propone una solución y en la sección 4 los resultados experimentales obtenidos.

2 Descripción del problema

Parte de la información que procesan los sistemas automáticos de recomendación web provienen de los algoritmos de búsqueda de reglas de asociación, que permiten extraer información de uso real del portal web. Entre otros algoritmos, Apriori contempla el soporte (porcentaje de sesiones que contienen cada página web visitada) y la confianza (porcentaje de sesiones que, teniendo el antecedente de la regla, también tienen su consecuente) para descubrir las reglas de asociación que contienen

los datos de uso del portal web. La idea esencial de Apriori consiste en generar un árbol L en que se guardarán las co-ocurrencias de páginas web en una misma sesión, usando inicialmente un conjunto de candidatos (C_k) para poder anotar el recuento de las páginas web que pueden ser de interés en cada nivel de L. Este árbol se crea siguiendo el siguiente esquema:

1. Generar el primer nivel del árbol extrayendo del repositorio de sesiones el número de veces que aparece cada una de las páginas visitadas (C_0).
2. Eliminar de C_0 aquellas páginas que no superen el umbral de soporte mínimo (L_0).
3. Generar el segundo nivel de candidatos (C_1) añadiendo a cada página de L_0 una rama por cada una de las restantes páginas del portal guardadas en L_0. De este modo no tendremos páginas candidatas que no superen el soporte mínimo fijado en el estudio.
4. Hacer sobre C_1 el recuento de los pares de páginas que aparecen en el repositorio de sesiones. Una vez anotadas todas las co-ocurrencias de páginas frecuentes en C_1 se eliminan todas aquellas que no superen el umbral fijado (L_1).
5. Se sigue el proceso $C_k \rightarrow L_k$ hasta que no se puedan generar nuevos candidatos.
6. Una vez tenemos L se extraen las reglas de asociación que superen el umbral de confianza mínima.

Esto da muy poco juego a las páginas nuevas que se incorporan al portal y a aquellas que son menos frecuentes, lo que no garantiza que sean de menor interés. Algunos estudios proponen incorporar al análisis la ponderación de las páginas web del portal de modo que se pueda forzar su incorporación al árbol L, sin embargo esto implica introducir (y mantener) información al análisis que no procede de los datos obtenidos a partir del uso del portarl web. Nosotros proponemos que de nuevo sean los datos procedentes de los usuarios quienes nos den esta información adicional. Imaginemos la siguiente información al construir L con Apriori con 10 sesiones, un soporte mínimo del 30% y una confianza mínima del 50%: la página A aparece en 9 sesiones, en 2 de las cuales también está la página B (que aparece en 3 ocasiones en total).

L_0	C_1
A(9)	B(2)
B(3)	

La regla A→B no se descubrirá porque el soporte observado (del 20%) hace que desaparezca de C_1 la página B al construir L_1 por lo que no será analizada como regla.

Veamos otra situación:

L_0	C_1
A(9)	B(3)
B(3)	

En este segundo caso sí que se mantiene B en L_1 pero la regla A→B no supera la confianza mínima (se observa un 33.3% de confianza) y no la registraremos.

Sin embargo cabe destacar que en el primer caso el 66.7% de veces que aparece B lo hace en una transacción en la que está A. Y en el segundo caso ocurre el 100% de

las veces. Con el planteamiento clásico de Apriori no obtenemos ninguna regla que tenga como antecedente la página A. Sin embargo hemos notado que en el uso del portal existe una relación entre la existencia de A y de B en la misma sesión:

> Si en una transacción está el ítem A no tenemos ni soporte ni confianza para decir que es probable que también aparezca el ítem B (de ahí que Apriori no nos advierta de nada) pero *si queremos sugerir B **éste** es uno de los mejores momentos para anunciarlo*.

Esta medida puede ser útil en otras ocasiones. Supongamos que tenemos 15 transacciones y un soporte mínimo del 20%, si observamos

L_0	L_1
A(10)	B(5)
	C(3)
B(10)	
C(3)	

En este caso la confianza de A→B es del 50%, muy superior al 30% de confianza que proporciona la regla A→C. Sin embargo la página B sólo se solicita conjuntamente con A el 50% de las veces que es visitada, frente al 100% que muestra C. Si sólo pudiéramos recomendar un enlace al usuario que ya ha solicitado la página A, el enfoque clásico de Apriori proporciona la página B como idónea, aunque se trate de un buen momento para sugerir la visita a la página C pues siempre se ha visitado conjuntamente con A.

Los umbrales de soporte y confianza mínimos surgen en el estudio de búsqueda de reglas de asociación debido a la gran cantidad de datos que deben manejarse. Aunque teóricamente se pueden encontrar todas las relaciones existentes en un gran repositorio de sesiones, en la práctica el número de relaciones es tan elevado que no puede ser tratado por un computador. Estos umbrales no permiten el análisis de las páginas que son visitadas con menor frecuencia (lo que ocurre a todas las páginas nuevas incorporadas al portal), lo que hace que sólo se trabaje con un porcentaje de las páginas del portal web y se ignoren por completo el resto de páginas.

En la siguiente sección se formaliza una nueva medida y un método que permite descubrir asociaciones interesantes entre las páginas de un portal web sin tener que renunciar a las relaciones expuestas en esta sección y con un coste computacional asumible por cualquier computador.

3 Reglas de oportunidad

El uso de soporte mínimo en la búsqueda de reglas de asociación, aunque es necesario para evitar que sea inabordable el estudio mediante computadores provoca la pérdida de información sobre un gran número de páginas del portal en estudio. La consecuencia más directa cuando se usa para alimentar un sistema de recomendación automático de enlaces es que sólo se sugiere visitar las páginas que ya son frecuentes con lo que su frecuencia crece y decrece la frecuencia relativa de las visitas a las páginas menos frecuentes. Si queremos ser capaces de recomendar cualquier página

del portal web a partir de sus datos de uso hemos de ser más flexibles con el uso del soporte mínimo.

Las *reglas de asociación* se generan sobre las páginas visitadas más frecuentemente en función de la confianza que ofrecen:

- Si al menos el 50% de los usuarios que visitan la página A también visitan en la misma sesión la página B se genera la regla A→B, sugiriendo al usuario que está visitando la página A que también debería visitar la página B.

Nosotros proponemos el uso de *reglas de oportunidad* sobre las páginas visitadas con menos frecuencia:

- Si al menos el 50% de los usuarios que visitan la página B (poco frecuente) también visitan en la misma sesión la página A se genera la regla A→B, sugiriendo al usuario que está visitando la página A que también debería visitar la página B. La regla inversa (B→A) sería una regla de asociación, pero no se genera pues no tiene soporte mínimo. El objetivo de las reglas de oportunidad es generar reglas cuyo consecuente no tenga soporte mínimo por lo que es poco probable que un usuario visite por sí mismo el consecuente y tendría poca utilidad si se planteara como antecedente.

Con esta nueva medida proponemos un nuevo algoritmo que es capaz de detectar este nuevo tipo de reglas incrementando considerablemente el porcentaje de ítems sobre el que tenemos información para el sistema de recomendación y sin incrementar apenas el uso de recursos del computador que realiza el análisis.

- En primer lugar debe ser capaz de recoger información de ítems que no superen el soporte mínimo. Si no consideráramos el soporte mínimo obtendríamos un árbol L extremadamente grande y con información irrelevante que desaparecería al obtener las reglas con confianza mínima. El soporte mínimo debe tener cierta flexibilidad.
- Las reglas de oportunidad con más de un antecedente no aportan mayor información al sistema de recomendación y generan muchos datos a almacenar por lo que se ignorarán. La explicación está en que si seguimos escribiendo en L_i, $i > 1$, la frecuencia del ítem "no frecuente" puede que la confianza de la regla clásica generada crezca pero nunca crecerá la oportunidad del ítem pues es una medida decreciente al avanzar por L.
- Debe informar de un nuevo tipo de reglas que pueden o no superar la confianza mínima.

Las modificaciones que proponemos sobre el algoritmo Apriori se reflejan en el siguiente algoritmo:

```
//En el primer nivel se recogen todos los ítems
while (quedan transacciones)
  lee_transaccion
  foreach (ítem en transacción)
    L_0[ítem]++
//Ya tenemos L_0 pues no consideramos soporte mínimo
```

```
//Generamos C_1
while (quedan transacciones)
  lee_transaccion
  foreach (2-itemset en transacción)
    L_0[item1]->C_1[item2]++

//Extraemos las reglas de oportunidad
foreach (item1 en L_0)
   foreach (item2 en L_0[item1]->C_1)
     if (L_0[item1]->C_1[item2] / L_0[item2] >= oportunidadMinima)
       añadir_regla_oportunidad(item1 → item2)

//Purgar L_0 y L_1 y seguir con el algoritmo clásico
...
end
```

4 Pruebas realizadas

Se han procesado datos sintéticos (T10I4D100K.dat y T40I10D100K.dat) y reales (BMS-POS.dat) para observar la incidencia de la obtención de reglas de oportunidad sobre datos de diversa procedencia, obteniendo resultados similares en todos los casos por lo que expondremos sólo los del primer repositorio.

El objetivo que tiene la introducción de reglas de oportunidad es el de obtener relaciones que cubran el mayor número de ítems computacionalmente posible. Queremos tener reglas que permitan, a un sistema de recomendación que se alimente de ellas, hacer inferencia sobre la idoneidad de sugerir cualquier ítem del repositorio. Según el enfoque clásico esto supone reducir el soporte mínimo a 0, con lo que aparecerá información sobre todos los ítems, pero los repositorios grandes o con un gran número de ítems distintos generan tal cantidad de información que desbordan la capacidad de las computadoras con que estamos trabajando. Sin embargo añadiendo las reglas de oportunidad sí podemos realizar el análisis.

El tiempo necesario para obtener las reglas de oportunidad es tan pequeño que no afecta al tiempo total necesario para buscar las reglas de asociación presentes en un repositorio. Si consideramos, sin embargo, la diferencia de tiempo necesaria para ejecutar el algoritmo con distintos soportes mínimos, para lograr información sobre un gran número de ítems del repositorio es notablemente más rápido trabajar con un soporte mínimo “grande” y añadir las reglas de oportunidad que trabajar sólo con reglas de asociación con un soporte mínimo más pequeño. Estas diferencias se han constatado al realizar los experimentos pero no se han tomado los datos de tiempo de ejecución pues no se buscaba dicha mejora en la experiencia realizada.

El menor de los repositorios puestos a prueba es T10I4D100K.dat. Contiene 870 ítems distintos en 100.000 transacciones de 10 ítems de promedio, con un total de 1.010.228 ítems. En la siguiente tabla se puede observar que al reducir el soporte mínimo con el enfoque clásico se obtiene información sobre un número mayor de ítems con un aumento exponencial del número de reglas encontradas, lo que dificulta enormemente su análisis.

Tabla 1. T10I4D100K con un 50% de confianza y oportunidad mínimas.

Soporte mínimo (%)	Reglas de asociación		Reglas de oportunidad		Ítems cubiertos por ambas (%)	Mejora aportada por las RO (%)
	Número de reglas	Ítems cubiertos (%)	Número de reglas	Ítems cubiertos (%)		
1,000	7	0,6	1.125	83,4	83,6	99,3
0,500	1.145	7,5	1.071	81,1	83,6	91,1
0,400	4.861	16,7	990	77,8	83,6	80,1
0,300	20.775	32,6	800	70,2	83,6	60,9
0,200	176.883	50,3	560	55,1	83,6	39,8
0,100	333.757	66,0	339	36,7	83,6	21,0
0,050	761.644	76,9	141	17,8	83,6	8,0
0,010	3.611.429	82,9	20	2,9	83,6	0,8
0,005	17.590.740	83,4	6	0,8	83,6	0,1
0,003	107.561.757	83,4	6	0,8	83,6	0,1
0,001	-	100,0	0	0,0	83,6	0,0

Para obtener información sobre un número mayor de ítems hemos de reducir los umbrales de confianza y oportunidad mínima. En la tabla 2 se observa este efecto y se comprueba que la mejora aportada por las reglas de oportunidad es ligeramente menor pero aún importante.

Tabla 2. T10I4D100K con un 25% de confianza y oportunidad mínimas.

Soporte mínimo (%)	Reglas de asociación		Reglas de oportunidad		Ítems cubiertos por ambas (%)	Mejora aportada por las RO (%)
	Número de reglas	Ítems cubiertos (%)	Número de reglas	Ítems cubiertos (%)		
1,000	17	1,4	2.913	99,9	99,9	98,6
0,500	1.535	21,0	2.710	99,0	99,9	78,9
0,400	5.923	34,0	2.482	97,7	99,9	65,9
0,300	23.199	53,8	2.004	93,3	99,9	46,1
0,200	189.205	72,8	1.328	81,4	99,9	27,2
0,100	354.998	88,7	596	53,2	99,9	11,2
0,050	808.183	95,5	217	24,8	99,9	4,4
0,010	3.895.609	99,4	25	3,3	99,9	0.5
0,005	19.642.125	99,8	10	1,5	99,9	0,1
0,003	126.446.840	99,8	10	1,5	99,9	0,1
0,001	-	99,9	0	0,0	99,9	0,0

5 Conclusiones y trabajo futuro

Los resultados obtenidos muestran que es posible obtener automáticamente información de uso de las páginas web nuevas o poco frecuentes. Al aplicarlo a un sistema de recomendación es de esperar que dé mayor "publicidad" a esas páginas y con ello que se incorporen antes al uso "frecuente" en el portal.

Descubrir patrones sobre ítems poco frecuentes puede ser útil en otras áreas como en la detección temprana de fraudes en tarjetas o seguros, en bioestadística o medicina para la inclusión automática de características poco frecuentes en el estudio de enfermedades…

Actualmente estamos estudiando una modificación del algoritmo que pueda decidir por sí mismo, en base a los datos analizados, qué niveles de soporte, confianza y oportunidad mínimos son los más adecuados para obtener reglas representativas de todos los ítems del repositorio.

6 Agradecimientos

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

References

1 Agrawal, R.; Imielinski, T. & Swami, A. Mining association rules between sets of items in large databases. Proc. of the 1993 ACM SIGMOD International Conference on Management of data, 207-216 (1993)

2 Kouris, I. N.; Makris, C. H. & Tsakalidis, A. K.: Using information retrieval techniques for supporting data mining. Data & Knowledge Engineering, Elsevier Science Publishers B. V., 52, 353-383 (2005)

3 Rozenberg, B. & Gudes, E.: Association rules mining in vertically partitioned databases. Data & Knowledge Engineering, Elsevier Science Publishers B. V., 59, 378-396 (2006)

4 Palshikar, G. K.; Kale, M. S. & Apte, M. M.: Association rules mining using heavy itemsets. Data & Knowledge Engineering, Elsevier Science Publishers B. V., 61, 93-113 (2007)

5 Tseng, M.-C. & Lin, W.-Y.: Efficient mining of generalized association rules with non-uniform minimum support. Data & Knowledge Engineering, 62, 41-64 (2007)

6 Lazcorreta, E.; Botella, F. & Fernández-Caballero, A.: Towards personalized recommendation by two-step modified Apriori data mining algorithm. Expert Systems with Applications, 35, 1422-1429 (2008)

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 Lee, W & Stolfo, S.J. & Mok, K.W.: *Mining audit data to build intrusion detection models.* Procs. of the 4th International Conference on Knowledge Discovery an Data Mining, (1998)

9 Han, J. & Fun, Y.: *Discovery of multi-level association rules from large databases.* Procs. of the International Conference on Very Large Data Bases, 420--431, (1995)

10 Liu, B. & Hsu, W. & Ma, Y.: *Mining association rules with multiple minimum supports*. Proc.. of the fifth ACM SICKDD International Conference on Knowledge Discovery and Data Mining, 337--341 (1999)

11 Hu, Y.-H. & Chen, Y.-L.: *Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism*. Decision Support Systems, Elsevier Science Publishers B. V., 42, 1--24 (2006)

12 Tseng, M.-C. & Lin, W.-Y.: Efficient *Mining of generalized association rules with non-uniform minimum support.* Data & Knowledge Engineering, 62, 41--64 (2007)