Visualización de Reglas de Asociación Un acercamiento comparativo

G. Filocamo W. Grandinetti S. Castro[1] S. Martig[1] C. Chesñevar[2]

Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
Av.Alem 1253 – B8000CPB Bahía Blanca – República Argentina
Tel/Fax: (+54) (291) 459 5135/5136 – Email: {smc,smartig,cic}@cs.uns.edu.ar
Palabras clave: Reglas de Asociación, Datamining, Visualización

Resumen

La minería de datos en búsqueda de *reglas de asociación* intenta encontrar relaciones de asociación o correlación entre grandes colecciones de datos. Si bien las reglas de asociación brindan una técnica que sintetiza una gran colección de datos, las mismas por sí solas pueden resultar poco informativas si no son visualizadas correctamente.

Este trabajo presenta distintos aspectos vinculados a la integración de reglas de asociación con técnicas de visualización. Se discuten las características de un software (**Miner.ar**) para obtención de reglas de asociación, y distintas alternativas para su integración con técnicas de visualización.

1 Introducción y motivaciones

El *descubrimiento de conocimiento a partir de bases de datos*, denominado genéricamente *datamining*, es un proceso de extracción no trivial de información implícita, previamente desconocida y potencialmente útil a partir de bases de datos. Dentro de las diversas técnicas existentes, las *reglas de asociación* constituyen una de las técnicas de datamining que intenta encontrar relaciones de asociación o correlación entre grandes colecciones de datos.

La obtención de reglas de asociación a partir de grandes bases de datos transaccionales resulta atractiva en diversas áreas. Algunos ejemplos de aplicación son la determinación de nuevas estrategias de marketing para organizaciones comerciales, detectección de fraudes financieros, determinación de la causa de enfermedades y utilización de patrones de información climática para predicción meteorológica [Toi96]. En tal sentido, cabe señalar que, si bien las reglas de asociación brindan una técnica que sintetiza una gran colección de datos, las mismas por sí solas pueden resultar poco informativas si no son presentadas al usuario final de manera correcta. Es aquí donde resulta importante contar con *técnicas de visualización* [CMS99, Chi00] apropiadas para resolver este problema.

Este trabajo describe los principales aspectos de una línea de investigación que ha comenzada a ser desarrollada en el ámbito del Departamento de Cs. e Ingeniería de Computación de la Universidad Nacional del Sur, que involucra la integración de reglas de asociación con técnicas de visualización. El trabajo se estructura como sigue: primeramente se describen los principales conceptos teóricos para obtención de las reglas de asociación. Luego se mencionan brevemente las características del software **Miner.ar** [FG02] para obtención de reglas de asociación, y finalmente se discuten los elementos más relevantes al momento de integrar dichas reglas con técnicas de visualización.

[1]Laboratorio de Visualización y Gráfica (VyGLab), UNS.
[2]Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA), UNS.

2 Reglas de Asociación en Bases de Datos

Seguidamente introduciremos las definiciones básicas necesarias para la caracterización de reglas de asociación. Como punto de partida se toma un conjunto de items $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$ y un conjunto $\mathcal{D}$ de transacciones, donde cada transacción $d \in \mathcal{D}$ tiene un identificador único y contiene un conjunto de items, también llamado *itemset*. Un itemset con k items es también llamado un *k-itemset*. El *soporte* de un itemset X, denotado como $\sigma(X)$, es el número de transacciones en las cuales X aparece como subconjunto. Un itemset es *maximal* si no es subconjunto de ningún otro itemset. Un itemset es *frecuente* si su soporte es superior al *mínimo soporte* (*min_sup*) especificado por el usuario.

Una **regla de asociación** es una implicación de la forma $X \Rightarrow Y$, $X \subset \mathcal{I}$, $Y \subset \mathcal{I}$ y $X \cap Y = \emptyset$. El *soporte* de una regla está dado por $\sigma(X \cup Y)$ y la *confianza* como $\sigma(X \cup Y)/\sigma(X)$ (i.e., la probabilidad condicional que una transacción contenga Y, dado que contiene X). Una regla es *fuerte* si su confianza es mayor a la mínima confianza específicada por el usuario (*min_conf*). NOTA: nótese que las reglas de asociación no son transitivas, en virtud de que $S \Rightarrow T$ y $T \Rightarrow U$ no implica necesariamente $S \Rightarrow U$.

El problema de extraer reglas de asociación de un conjunto de transacciones $\mathcal{D}$ consiste en generar todas las reglas de asociación que tengan un soporte y confianza mayor que el mínimo especificado por el usuario (*min_sup* y *min_conf* respectivamente). Esta tarea puede ser descompuesta en dos partes:

- *Buscar los itemset frecuentes.* Dados m ítems existen potencialmente 2^m itemset frecuentes, por lo que su computación por fuerza bruta es inviable. Se requieren algoritmos especializados para esta tarea.
- *Generar todas las reglas de asociación fuertes.* Estas serán reglas de la forma $X \backslash Y \Rightarrow Y$ donde $Y \subset X$, para todos los itemsets X frecuentes que superen el mínimo de confianza.

La obtención de los itemset frecuentes es en consecuencia un aspecto clave en la determinación de reglas de asociación. Existen diversas técnicas, destacándose las siguientes:

Apriori [AS94] Ampliamente utilizado por su simplicidad. Utilizando el principio anterior se generan los itemset de tamaño k candidatos a partir de los itemset frecuentes previos.

FP-Growth [HPY00] Es un método más reciente y supera al anterior. Aplica la técnica de *dividir y conquistar* y con ello reduce el espacio de búsqueda, transformando el problema de buscar largos fragmentos de patrones en buscar fragmentos más cortos y luego realizar la concatenación de sufijos construyendo un árbol y trabajando sobre el mismo.

2.1 El software Miner.ar. Características

El sistema **Miner.ar** [FG02] es una herramienta desarrollada en Java para extracción de reglas de asociación. Provee una interface gráfica que permite tomar una base de datos arbitraria, seleccionar atributos específicos y extraer reglas de asociación a partir de parámetros definidos por el usuario. El método de extracción usado es *FP-Growth*. Se realizaron experimentos de su performance a partir de una base de datos con información administrativa de la Universidad Nacional del Sur (más de 12.000 transacciones).[3]

[3] Para mayores detalles consultar [FG02].

3 Visualización de Reglas de Asociación

Las *reglas de asociación* constituyen uno de los patrones más importantes que pueden ser descubiertos utilizando datamining. Su exploración es una de las tareas centrales en el proceso de la minería de datos. Se ha invertido mucho tiempo y esfuerzo de investigación en el estudio de las mismas, principalmente en lo que hace a su generación o descubrimiento de una manera eficiente. La inclusión de algoritmos que permitan su descubrimiento de manera automática en la mayoría, sino en todas, las herramientas de datamining es un indicativo de su importancia. Centramos nuestra atención en dos aspectos de la manipulación de las reglas de asociación que aún no han sido totalmente resueltos:

- La gran cantidad de reglas generadas.
- La dificultad en el entendimiento de las reglas de asociación.

Realizaremos nuestro abordaje a los problemas planteados desde la Visualización de Información.

3.1 Visualización: Conceptos preliminares

La *Visualización de Información* favorece el entendimiento y el análisis de datos abstractos mediante el uso de Computación Gráfica Interactiva y de Técnicas de Visualización. Los datos abstractos no son inherentemente geométricos y presentan desafíos a los investigadores en Visualización porque evidentemente no es obvio poner los datos abstractos en formas visuales efectivas. La Visualización de Información es la representación gráfica adecuada tanto de los datos con parámetros múltiples como de las tendencias y las relaciones subyacentes que existen entre ellos. Su propósito no es la creación de las imágenes en sí mismas sino la asimilación rápida de información o monitoreo de grandes cantidades de datos. Este medio es promisorio fundamentalmente por varias razones:

- Acrecienta los recursos del humano en la forma de procesamiento perceptual expandiendo su memoria de trabajo.
- Favorece el reconocimiento de patrones.
- Permite el uso de inferencia y monitoreo perceptual.
- El medio mismo es manipulable e interactivo.

Un aspecto fundamental en las técnicas de Visualización de Información es el de la interacción, pues ésta es central para el proceso exploratorio del espacio de información, siendo la generadora de un proceso de retroalimentación que potencia la tarea. En un sentido amplio, podemos atacar el problema de visualizar grandes volúmenes de información desde distintas perspectivas:

- En uno de los extremos tenemos la visualización de los datos propiamente dichos, tratando de mostrar visualmente todos los items de datos, o su mayoría, y proveer técnicas que favorezcan su exploración. La principal limitación de estas técnicas es que podemos mostrar simultáneamente un número limitado de ítems de datos.

- Una alternativa es mostrar visualmente algún tipo de metainformación (por ej. algún tipo de resultado producido por datamining como son las reglas de asociación) y favorecer la exploración visual de ésta en un segundo nivel de abstracción o en una etapa de sintonía fina. Esta estrategia, si bien manipula conjuntos de datos de cardinalidad inferior a la del conjunto original, debe permitir visualizar los conjuntos de datos generados por los algoritmos de datamining que suelen ser de gran tamaño y difíciles de entender para los usuarios.

3.2 Aspectos relevantes a visualizar en Reglas de Asociación

Como se vio anteriormente, las reglas de asociación de la forma $X \Rightarrow Y$, son interpretadas como: *"Si en una transacción ocurre X, entonces es frecuente que en la misma transacción también ocurra Y"*. En este contexto, la noción de "frecuente" es cuantificada por dos medidas (el soporte y la confianza), las cuales reflejan la utilidad y la certeza de la regla repectivamente. Es indudable que la manera en que se visualicen las reglas deberá proveer información sobre:

- Los valores que componen X
- Los valores que componen Y
- La confianza de cada regla
- El soporte que tiene cada regla

Previamente se estableció que el problema de las reglas de asociación ofrecía por lo menos dos frentes: **cantidad de las reglas** y el **entendimiento de las mismas**. En cuanto a la cantidad de las reglas generadas por los algoritmos, se ha realizado mucho trabajo para filtrar y generar conjuntos de reglas de asociación fuertes. El problema surge del hecho de que las reglas desechadas pueden aportar al usuario información que le sería de utilidad para entender las reglas fuertes que quedaron en el conjunto. Lo anterior no significa que los algoritmos no deban realizar un filtrado de las reglas, todo lo contrario, lo que sí se debe considerar es la manera de ofrecer información adicional y herramientas que le permitan al usuario comprender la estructura subyacente de las reglas de asociación.

4 Línea de investigación a desarrollar

En una primer instancia se realizará un relevamiento de las distintas técnicas utilizadas en distintas herramientas disponibles para los fines propuestos. Se evaluarán las mismas con respecto a diferentes perfiles de usuario desde el punto de vista de la efectividad y la facilidad de uso de las mismas.

Se realizará en particular un análisis de los acercamientos utilizados en los principales paquetes de software para datamining ya existentes (ej. DBMiner). Se contrastarán las técnicas usadas en estos paquetes con aquellas técnicas que han sido exploradas e implementadas en el ámbito del VyGLab.

En una etapa posterior se tratará de combinar el software **Miner.ar** con las herramientas visuales para potenciar el software existente, a partir de la incorporación de diferentes técnicas de visualización.

5 Trabajos relacionados. Conclusiones

La visualización de reglas de asociación ha sido explorada en distintas variantes. Entre algunas de las muchas alternativas propuestas pueden mencionarse el uso de un *ball graph* (grafo de bolas) que interrelaciona distintas reglas de asociación, la utilización de una 'grilla' de reglas[4] y propuestas basadas en 'Mosaic plots' [HSW00]. También se ha estudiado el análisis visual de reglas de asociación a lo largo de un período de tiempo [ZL00].

La extracción de reglas de asociación a partir de grandes bases de datos transaccionales resulta atractiva en diversas áreas, y su aplicabilidad está condicionada en gran medida a la posibilidad de visualizarlas adecuadamente.

Si bien se han desarrollado diferentes abordajes en la visualización de reglas de asociación, entendemos que los mismos carecen muchas veces de un enfoque unificador y obedecen a intuiciones de los diseñadores de software. Consideramos que dicho enfoque puede ser provisto y sistematizado a partir de las técnicas de visualización existentes, muchas de ellas desarrolladas a nivel prototípico en el VyGLab. En esta dirección se desarrollarán las etapas de investigación futuras.

Referencias

[AS94] Agrawal, R., and Srikant, R. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference. Santiago, Chile* (1994).

[Chi00] Chi, E. A taxonomy of visualization techniques using the data state reference model. In *IEEE Visualization 2000, Actas en CD-Rom* (2000).

[CMS99] Card, S., Mackinlay, J., and Shneiderman, B. *Readings in Information Visualization - Using Vision to Think*. 1999.

[FG02] Filocamo, G. R., and Grandinetti, W. Reglas de asociación para datamining – teoría y aplicaciones. *Tesis de Licenciatura – Dep. de Cs. e Ing. de la Computación – Universidad Nacional del Sur* (2002).

[HPY00] Han, J., Pei, J., and Yin, Y. Mining frequent patterns without candidate generation. In *Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data, Dallas, TX* (May 2000).

[HSW00] Hofmann, H., Siebes, A., and Wilhelm, A. Visualizing association rules with interactive mosaic plots. In *KDD 2000* (2000), ACM.

[Toi96] Toivonen, H. *Discovery of Frequent Patterns in Large Data Collections*. PhD thesis, University of Helsinki, Finland, 1996.

[Tuf97] Tufte, E. *Visual Explanations: Images and Quantities, Evidence and Narrative*. 1997.

[WB95] Wong, P. C., and Bergeron, R. D. 30 years of multidimensional multivariate visualization. In *Proc. Workshop on Scientific Visualization* (1995), IEEE Computer Society Press.

[ZL00] Zhao, K., and Liu, B. Visual analysis of the behavior of discovered rules. In *KDD 2000* (2000), ACM.

[4] Estas técnicas en particular han sido utilizadas en el software comercial DBMiner.