Análisis de canastas de compras utilizando árboles de mínima distancia
(*Market basket analysis using minimum spanning trees*)

Mauricio A. Valle, Gonzalo A. Ruz, Rodrigo Morrás

Resumen

Se presenta una metodología de análisis de canasta de compras basada en árboles de mínima distancia (MST por sus siglas en inglés). Debido a la amplia variedad de productos en una tienda y a la heterogeneidad del comportamiento de compras de los consumidores, el análisis de canastas de compras es muy complejo. La metodología propuesta simplifica significativamente el proceso de encontrar conjuntos de productos que tienen alta co-ocurrencia en la canasta de compra de los consumidores. Es decir, productos que se llevan de manera conjunta. El resultado del análisis es una representación visual de red que conecta todos los productos con una alta correlación entre sí, y elimina posibles conecciones de productos con una correlación baja. Esta solución resulta ser más práctica que la tradicional reglas de asociación utilizadas para el análisis de canastas de compras.

Keywords: canastas de compra, redes, árboles de mínima distancia, reglas de asociación.

Introducción

Como administrador de una cadena de retail, un desafío permanente consiste determinar estrategias de promoción y de ubicación (*layout*) que permitan maximizar las ventas. Por ejemplo, ¿qué productos debieran promocionarse en forma conjunta?, ¿cuál es la major disposición de productos en una góndola para aumentar la posibilidad de que sean parte de una canasta de compra? ¿qué productos son complementarios dentro de una gama de productos de una misma categoría?

Las respuestas para este tipo de interrogantes pueden encontrarse en el análisis de canastas de compras o *market basket analysis* (MBA). El MBA es un método de minería de datos enfocado a encontrar encontrar patrones de compras de los consumidores en bases de datos transaccionales (Linoff y Berry, 2011). Una base de dato transaccional contiene al menos, las identidades o identificadores de los consumidores, el conjunto de productos (o ítemes) comprados (la cual constituye "la canasta de compra"), y la cantidad de cada ítem comprado. De esta forma, el MBA tiene como objetivo encontrar asociaciones de productos, es decir, identificar qué productos o suelen llevarse juntos y cuáles no.

La harramienta tradicional de minería de datos, con la cual el MBA ha campo de diversas aplicaciones en marketing-retail, ha sido las reglas de asociación (o *association rules mining* en inglés). Las reglas de asociación (Agrawal y Srikant, 1994) o ARs, permiten encontrar declaraciones del tipo, "cuando se compran fideos, entonces también llevan salsa de tomate". Es decir, se identifica la consecuencia: "la salsa de tomate", o el lado izquierdo de la regla, y el antecedente: "la compra de fideos", o el lado derecho de la regla. La popularidad de las reglas de asociación radica en la forma explícita en que se establece la asociación entre productos (antecedente y consecuencia), por lo que es fácil de interpreter para tomar acciones que permitan, por ejemplo, cambiar o modificar ofertas de productos amarrados, o cambiar promociones dirigidas aumentar la fidelidad del cliente.

La mayor desventaja de las reglas de asociación es que en la práctica, el algoritmo suele encontrar un número de reglas tan numerosas, que la actividad analizar aquellas que sean interés para el analista o para el administrador, suele ser una actividad no trivial y de alto consumo de tiempo (Klementtinen, et al, 1994). Aún cuando hay diversos índices disponibles para evaluar la calidad de las reglas encontradas, al ordenarlas por distintas medidas, el ranking de reglas ordanadas según su calidad varía, y por lo tanto, no es possible actualmente tener una medida objetiva para determinar qué reglas son mejores que otras en su capacidad de explicar el comportamiento de compra.

Este estudio, propone una alternativa complementaria a las reglas de asociación, la cual permite de manera visual y rápida, encontrar asociaciones de productos en términos de una "red de productos" (Reader y Chawla, 2009, 2011) basada en árboles de mínima distancia (o en inglés *minimum spanning trees*, MST). En otras palabras, la topología de la red por sí sola permite identificar asociaciones entre pares de productos, y en consecuencia, reducir el espacio de búsqueda de reglas de asociación a sólo aquellas que revela la red.

El presente estudio presenta una metodología basada en MST, que permite identificar asociaciones de productos significativas en una muestra de canasta de productos, de manera intuitiva y clara, descartando potenciales reglas espúreas y dejando sólo aquellas tienen alto nivel de dependencia.

Conceptos y definición del problema

Reglas de asociación:
Una canasta de datos (*Basket data*) representa el conjunto de k ítemes o productos disponibles para el consumidor. Se define como $I = \{i_1, \ldots, i_k\}$. Las reglas de asociación intentan encontrar dependencia entre ítemes que forman parte de canastas de compras de los clientes y que se encuentra en una base de datos transaccional D.
Se define soporte o *support* el grado de popularidad de un ítem y se mide como la proporción de transacciones en la cual el ítem aparece. En otras palabras, representa la probabilidad de que el ítem i_k esté presente. En otras palabras, $Suport\{i_k\} = P(i_k)$. La confienza o *confidence* es cuán probable que un ítem i sea llevado (consecuencia) cuando se también se compra el ítem k (antecedente), y se mide como la proporción de transacciones con el ítem i_j en la que el ítem i_k también aparece, es decir, $Confidence\{i_j \rightarrow i_k\} = P(i_j \mid i_k)$. Finalmente, la elevación o Lift mide cuán probable es que el ítem i_j sea comprado cuando también se compra i_k, controlando por la popularidad del ítem i_j. $Lift\{i_j \rightarrow i_k\} = P(i_j \mid i_k)/P(i_k)$. Si la elevación de una regla es mayor a 1, entonces la ocurrencia del antecedente y de la consecuencia son dependientes una de otra, y por lo tanto, la regla es potencialmente útil en predecir la consecuencia en futuras transacciones.

Específicamente, se dice que $i_1 \rightarrow i_2$, cuando:

1. i_1 y i_2 ocurren juntos al menos un s% de las n canastas de compras. Las reglas encontradas deben tener al menos este nivel de soporte,
2. y de todas las canastas que contienen i_1, al menos un c% también contienen i_2. Las reglas encontradas deben tener al menos este nivel de confianza.

Los parámetros s% y c% son dados por el analista. Bajo estas dos condiciones, todas aquellas reglas encontradas las cumplan, serán objeto de análisis. Tal como se indicaba anteriormente, la principal limitación de las ARs es que para una base de datos transaccional, la cantidad de

reglas que se generan cumpliendo las dos condiciones previas pueden ser de cientos o miles, las cuales, aún siendo ordenadas por Lift u otra medida de calidad, es difícil analizarlas todas o encontrar reglas que sean especialmente interesantes para el administrador del retail.

Redes de productos:
Las redes de productos se pueden construir a partir de una lista de transacciones. Cada nodo representa un ítem y los lazos que conectan cada vértice de la red representan si el par de ítemes fueron comprados en una misma transacción (Reader and Chawla, 2011). Esa es la descripción intuitiva de la red de productos, no obstante, de manera similar a lo que ocurre en muchos otros fenómenos representados por redes, se observa que hay un conjunto pequeño de nodos o vertices que tienen una alta conectividad (a otros nodos), es decir un alto grado (*degree*), mientras que muchos nodos tienen una muy baja conectividad (bajo grado). Esto significa que la distribución del grado de la red es altamente sesgada a la derecha, dando a lugar las típicas redes libres de escala (Barabási, 2009).
El alto grado de sesgo de la distribución del grado de la red de productos impone problemas prácticos porque para para redes de productos, la existencia de un lazo no significa necesariamente que dos productos tengan presencia en cada canasta de compra. Adicionalmente, es fácil imaginar que tratándose de cientos de productos existentes en un almacén, la cantidad potencial de lazos es enorme. Por ejemplo, una red con solo 200 ítemes, existen potencialmente 4,950 lazos, lo cual hace muy difícil encontrar patrones de compra y en consecuencia, poco práctico.
Para superar este problema, se han planteado diversas alternativas. Por ejemplo, el enfoque de detección de comunidades de productos es ampliamente conocido en redes y se ha utilizado en diversos campos de aplicación, especialmente en genética (Blondel et al, 2008; Clauset et al, 2004; Newman, 2006). En redes de productos, se ha utilizado la "densidad de información" como parámetro para encontrar comunidades de productos (Reader and Chawla, 2011), la cual mide el grado de información presente en una comunidad. Así, mientras más información tenga un grupo de productos, dicho grupo sera más susceptible de llegar a convertirse en una comunidad.
Otra alternativa aparentemente efectiva y de mayor simplicidad, consiste en al aplicación de umbrales al los grados de la red (Videla-Caviares and Ríos, 2014). Es decir, aquellos nodos (ítemes) con un grados bajo un valor determinado, se remueven de la red, dejando solo aquellos nodos que presentan mayor conectividad. Luego, es possible encontrar y describir zonas de productos con una fuerte relación entre ellos.
Si bien estos métodos de análisis de canastas de compras utilizando redes de productos logran capturar valiosa información de las relaciones que existen entre una amplia gama de ítemes, carecen aún de una representación visual y práctica para ser utilizada en terreno y para lograr tomar decisiones in situ. Adicionalmente, los enfoques utilizados usualmente son variaciones de otras soluciones aplicadas en otras disciplinas, mientras que este estudio propone una solución ad-hoc al problema de las canastas de compras, considerando que, en escencia, dicho problema es uno de identificación de correlaciones (Brin, Motwani y Silverstein, 1997).

Árboles de mínima distancia (MST):
Los MST son un tipo especial de grafo en que los todos los nodos quedan conectados sin formar bucles. De este modo, los n vertices de la red, tendran a lo más $n - 1$ lazos. Cada nodo A diferencia de una red de productos en que el lazo representa la presencia o ausencia de productos en una misma canasta de compra, en el MST, los lazos representan el grado de correlacion entre los ítemes, la cual es apropiadamente transformada a una medida de distancia.

Los MST han sido utilizados para estudiar las interacciones existentes en el mercado de divisas. Por ejemplo se han utilizado para estudiar el comportamiento sistémico de los precios de las acciones del mercado de valores de Estados Unidos (Bonanno et al, 2003). El mercado de divisas extranjero ha sido estudiado como una red MST, encontrando agrupaciones de monedas altamente relacionadas entre sí. Recientemente, métodos para filtrar información han sido utilizadas para modelar correlaciones entre los precios de acciones del DAX alemán, creando visualizaciones de redes basadas en correlación. (Birch, Pantelous and Soramäki, 2016; Onnela, Kaski and Kertész, 2004). Estos son solo algunos ejemplos del potencial que tienen los MST para ganar mayor entendimiento del comporamiento de mercados. En este proyecto, el fenómeno consiste en las decisiones de compra de consumidores. Evidentemente, al tratarse de una gran cantidad de nodos (productos), se podría pensar en la necesidad de aplicar algun tipo de filtro con el objeto de retirar enlaces con baja correlacion (alta distancia entre nodos). Sin embargo, una ventaja del MST es que este tipo de redes muestran solo los nodos (productos) mas importantes generando una red compacta. De hecho, el MST permite encontrar la organización jerárquica de los nodos (Mantegna, 1999) de manera tal, que es posible descubrir grupos de categorías o productos homogéneos con respecto a las preferencias de los consumidores en el conjunto del total de productos ofrecidos por el supermercado.
Dado que la red se basa en la correlación (y no meramente en la presencia de productos en la misma canasta), este estudio se basa en el fundamento de que las comunidades de productos que se forman en la red, estan expuestos a menos ruido y en consecuencia, representan de manera mas limpia las agrupaciones de ítemes que estan íntimamente relacionados entre si. De esta forma, se obtiene una topología que representa en forma resumida una gran cantidad de transacciones a un nivel que sea manejable o interpretable por un administrador.

Metodología

La metodología propuesta de análisis de canastas de compra se resumen en los siguientes 4 pasos:

1. Obtener matriz de correlaciones entre productos.
2. Transformar las correlaciones en una medida de distancia.
3. Obtener el MST
4. Detectar lazos significativos y comunalidades de productos.

El primer paso consiste en obtener una medida de asosiación entre los productos de la canasta. La base de datos transaccional mantiene registro de los productos que se llavan en cada transacción, de modo que la presencia o no presencia del producto en una canasta se representa como una variable binaria. Así, se calcula el índice de correlacion phi (Knobbe and Adriaans, 1996), que representa el grado de asociación entre variables binarias. Para n productos, el resultados es una matriz simétrica de nxn. Al igual que en la correlación de Pearson, la correlación phi toma valores entre -1 y 1.
Luego, es necesario transformar dichas correlaciones en una métrica de distancia, de forma tal que cuando la correlación sea igual a 1, la distancia es nula. Una transformación que cumple con dicho criterio es:

$$d_{ij} = \mathrm{sqrt}(2(1-\mathrm{phi}_{ij})) \qquad [1]$$

donde phi_{ij} es el coeficiente phi entre ítem *i* y *j*, y sqrt es la raíz cuadrada. La matriz de correlaciones es ahora transformada de acuerdo a la expression [1]. La simetría se mantiene, es decir, $d_{ij} = d_{ji}$, con valores nulos en la diagonal.
El tercer paso, consiste en identificar el MST a partir de la matriz de distancia. Esto se lleva a cabo con el algoritmo de Prim o de Kruskal (Graham y Hell, 1985). Finalmente, se aplica un proceso de agrupamiento jerárquico desde el cual es posible detectar aglomeraciones o grupos de productos. El fundamento de esta agrupación se base en el hecho de un producto j en el MST queda a una distancia de otro producto i, de tal manera que todo el conjunto de productos que queden a una distancia muy pequeña entre ellas, pero lejanas a otro grupo, son cadidatos de formar un cluster de productos. La significancia práctica de los clusters encontrados son relevantes porque revelan el conjunto de ítemes que contienen productos complementarios, es decir que suelen llevarse en una misma transacción.

Como ejemplo de aplicación, se ha tomado una base de datos transaccional real que contiene 9835 transacciones y 169 ítemes de un almacén típico correspondientes a 30 días (Hahsler, Hornik, y Reutterer, 2006). Para tener una idea de la variabilidad y complejidad de las transacciones, una muestra de 80 ítemes se han extraído para graficar un mapa de calor (*heatmap*) de las transacciones (Figura 1). El color oscuro representa la presencia del ítem en la transacción, mientras que color blanco representa la no presencia.

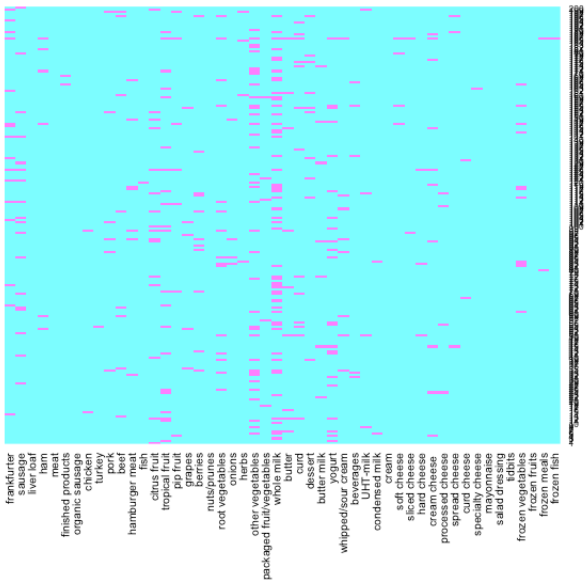

Figura 1 – Heatmap de una muestra aleatoria de 200 transacciones y de 80 ítemes.

La observación de la Figura 1 pone de manifiesto la complejidad del análisis de las canastas de compras debido a la variedad de productos y particularmente por la heterogeneidad de las composiciones de cada canasta. La metodología propuesta ayuda precisamente a reducir la complejidad de los datos para poder descibrir patrones de compra significativos.
Un análisis simple de frecuencia de ítemes presentes en la canasta de compra revela que los productos más comprados con lechen entera (whole milk), otros vegetales (other vegetables), rollos de pan (rolls buns) y bebida soda (soda). Ver Figura 2.

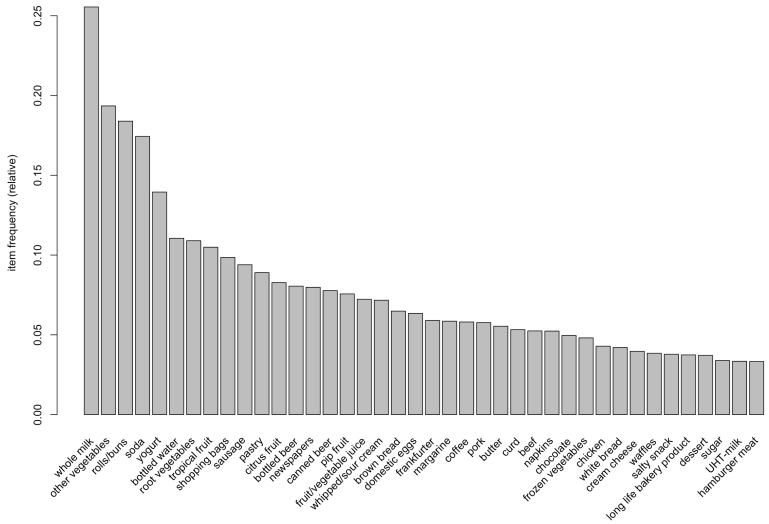

Figura 2 – Frecuencia de ítemes para la base de datos transaccional.

La Figura 2 revela claramente aquellos productos que tienen mayor rotación en el almacén, lo cual puede ser de utilidad para establecer estrategias a productos con mayor volúmen de ventas. Sin embargo, esta información básica no dice nada a cerca de la relación que tienen estos productos con otros presentes en una misma canasta. Por ejemplo, podría ser que un producto con baja frecuencia de compra se lleve con casi total seguridad con algún otro producto complementario. Desde el punto de vista de la estrategia promocional, la rentabilidad de estos productos puede ser mucho mayor que la de otro con mayor frecuencia de compra.

Resultados

La Figura 3 muestra el resultado de la aplicación de la metodología. El MST encontrado refleja al menos 13 ramas de productos que emanan de la espina principal, lo cual significa potenciales candidatos de clusters de ítemes. Por otro lado, mientras el lazo que conecta cada nodo sea de menor tamaño, indica que los ítemes se llevan al mismo tiempo. Por ejemplo, es fácil ver que ketchup y mostaza (*mustard*) están muy juntos, por lo que ambos ítemes tienden a comprarse al mismo tiempo. Por ejemplo, al extremo derecho del MST, se observa que azúcar (*sugar*), café (*coffee*), sal (*salt*), mermelada (*jam*), leche condensada (*condensed milk*) y harina (*flour*) son productos que suelen estar presentes en una misma canasta de compra. Por ejemplo, en el extremo izquierdo del MST se observa que existe una alta propensión a llevar licores de la mano de vino y cerveza.

Por otro lado, se observa también productos con alto nivel de correlación pero con poca nivel de complement. Por ejemplo, bolsas de compra (*shopping bags*) y las salchicas (*sausage*) se llevan en una misma compra. La administración del almacén debiera intentar responser a este tipo de comportamientos, pudiéndos ser una respuesta específica de los clientes a un layout particular o algún otro efecto externo.

Finalmente, se reconoce en el MST productos que topológicamente aparecen aislados, como por ejemplo, el queso en rodajas (*sliced cheese*) o el jugo de frutas (*fruit/vegetable juice*). De esta manera, una estrategia de promoción de productos amarrados tendría más éxito si se considera el pack "ketchup y mostaza" que un pack que incluya el "jugo de frutas y vegetales". En otro sentido la salchicha especial (*frankfurter*) tiene un alto nivel de volúmen de ventas. Este ítem está dentro de los primeros 15 productos con mayor rotación, no obstante este ítem aparece de manera aislada en el MST, de modo que enfocar una campaña para

aumentar el volumen de ventas de otro producto atándolo con este tipo de salchicha no sería conveniente.

Para demostrar la simplicidad de la metodología propuesta en contraste con las reglas de asociación, se llevó a cabo un proceso de descubrimiento de reglas utilizando el paquete Arules (Hahsler et al, 2014) para R (Team, 2014).

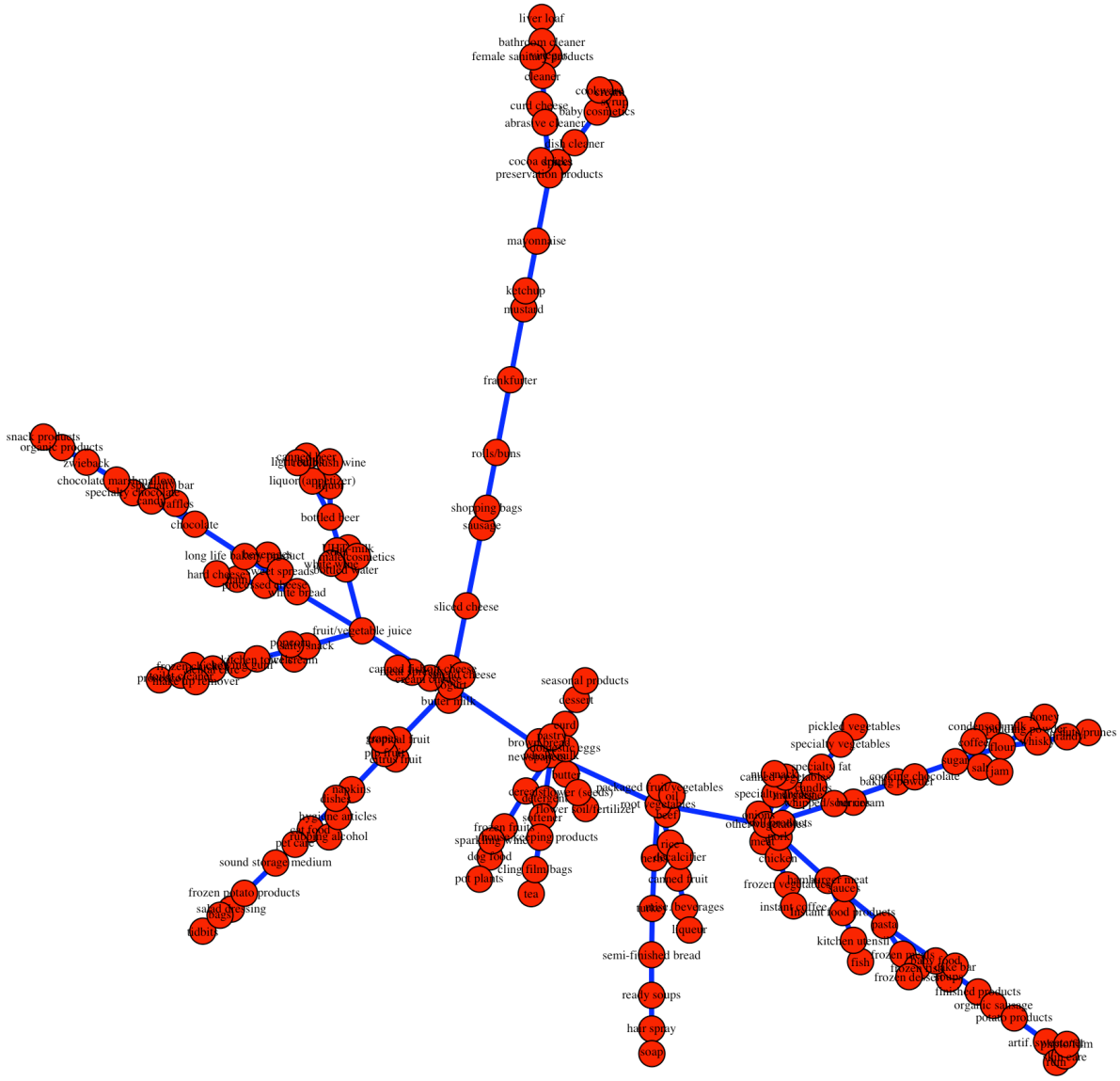

Figura 3 – Representación del MST completo para la base de datos transaccional.

Utilizando un soporte y confianza mínimo de 0.001 y 0.02 respectivamente, se encontraron 180499 reglas de asociación. Para tener una idea de la complejidad de la actividad del análisis de las reglas de asociación, la Figura 4 muestra un gráfico de dispersión de las reglas con *Lift* superior a 1, las cuales suman un total de 49247. Es decir, si bien el número de reglas se ha

reducido significativamente, el número de reglas, aún es muy elevado como para realizar un análisis práctico y recomendar estrategias de promoción o marketing.
Para observar con mayor detalle el conjunto de reglas de asociación, la Tabla 1 presenta las 10 reglas (de las 49247) con el mayor nivel de *Lift*. Por ejemplo, el antecedente de primera regla es botella de cerveza y vino , mientras que la consecuencia es licor. La interpretación de esta regla es que cuando se lleva botella de cerveza y vino, entonces también se compra Licor. Esta regla de deduce fácilmente del MST (Figura 3, costado izquerdo superior de la red). Por ejemplo la regla {processed cheese,white bread} => {ham} también es fácil detectarla en el MST. De hecho, las reglas con el mayor nivel de *Lift* se encontrarán en el MST, mientras que aquellas con un Lift de menor intensidad simplemente no quedan representadas en la red. Esto pone de manifiesto la utilidad de la metodología basada en MST.

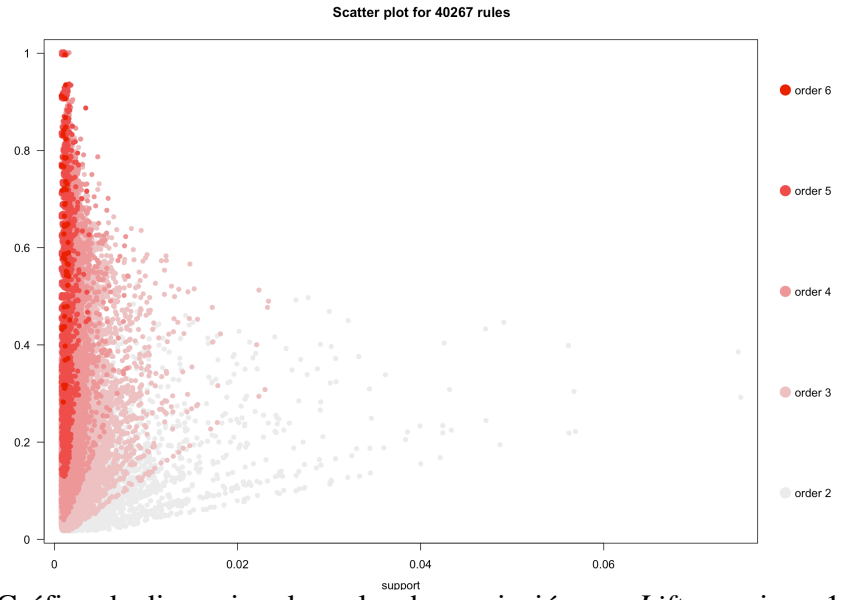

Figura 4 – Gráfico de dispersion de reglas de asociación con *Lift* superior a 1. Al costado derecho de la gráfica se muestra el nivel de *Lift* según la intensidad del color. El eje vertical indica nivel de confianza (*confidence*) de las reglas.

Tabla 1 – Muestra de las primeras 10 reglas encontradas con mayor nivel de *Lift*.

	lhs	rhs	support	confidence	lift
[1]	{bottled beer,red/blush wine}	=> {liquor}	0.001931876	0.3958333	35.71579
[2]	{hamburger meat,soda}	=> {Instant food products}	0.001220132	0.2105263	26.20919
[3]	{ham,white bread}	=> {processed cheese}	0.001931876	0.3800000	22.92822
[4]	{root vegetables,other vegetables,whole milk,yogurt}	=> {rice}	0.001321810	0.1688312	22.13939
[5]	{bottled beer,liquor}	=> {red/blush wine}	0.001931876	0.4130435	21.49356
[6]	{Instant food products,soda}	=> {hamburger meat}	0.001220132	0.6315789	18.99565
[7]	{curd,sugar}	=> {flour}	0.001118454	0.3235294	18.60767
[8]	{soda,salty snack}	=> {popcorn}	0.001220132	0.1304348	18.06797
[9]	{sugar,baking powder}	=> {flour}	0.001016777	0.3125000	17.97332
[10]	{processed cheese,white bread}	=> {ham}	0.001931876	0.4634146	17.80345

Conclusiones

El MST equivale a una representación gráfica del rol que juegan los principales ítemes de la canasta de compras de una tienda de almacén, y así también los ítemes que tienen un papel

periférico en influir la canasta de compra. Los nodos que poseen un grado mayor que dos se consideran como más importantes debido a su mayor influencia en la canasta de compras. Tal como se observó en la aplicación, fue posible encontrar rápidamente aquellos productos con una alta propensión a ser incluidos en forma conjunta en la canasta de compras. De esta manera, esta herramienta promete ser una excelente alternativa para proponer estrategias de márketing in-situ.

Una siguiente etapa en el estudio de análisis de canasta de compras es aplicar la metodología a una base de datos transaccional de empresas latinoamericanas. Particularmente, se dispone de información de una cadena de supermercados en distintos locales geográficos de la capital de Chile. Se espera que el MST sea capaz de identificar conductas de compras distintivas entre las distintas sucursales de la misma cadena de supermercados, lo cual permite establcer estrategias de ofertas y promoción a nivel "local" y no necesariamente a nivel nacional.

Referencias

Agrawal, R., & Srikant, R. (1994). *Fast algorithm for mining association rules in large database*. Research report RJ 9839, IBM Almaden Research Center, Santiago, Chile.

Barabási, A. L. (2009). Scale-free networks: a decade and beyond. *Science*, 325(5939), 412-413.

Birch, J., Pantelous, A. A., & Soramäki, K. (2016). Analysis of correlation based networks representing DAX 30 stock price returns. *Computational Economics*, *47*(4), 501-525.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Bonanno, G., Caldarelli, G., Lillo, F. & Mantegna, R.N. (2003). Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review E*, 68(4):046130, 2003.

Brin, S., Motwani, R., & Silverstein, C. (1997, June). Beyond market baskets: Generalizing association rules to correlations. In *ACM Sigmod Record* (Vol. 26, No. 2, pp. 265-276). ACM.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*(6), 066111.

Graham, R. L., & Hell, P. (1985). On the history of the minimum spanning tree problem. *Annals of the History of Computing*, *7*(1), 43-57.

Hahsler, M., Hornik, K., & Reutterer, T. (2006). Implications of probabilistic data modeling for mining association rules. In *From Data and Information Analysis to Knowledge Engineering* (pp. 598-605). Springer Berlin Heidelberg.

Hahsler, M., Buchta, C., Gruen, B., Hornik, K., & Hahsler, M. M. (2014). Package 'arules'.

Kwapień, J., Gworek, S., Drożdż, S., & Górski, A. (2009). Analysis of a network structure of the foreign currency exchange market. *Journal of Economic Interaction and Coordination*, *4*(1), 55.

Knobbe, A. J., & Adriaans, P. W. (1996, August). Analysing Binary Associations. In *KDD* , Vol. 96, p. 311.

Klementtinen, M., Mannila, H., Ronkainen, P., Toivonen, H., & Verkamo, A. (1994). *Finding interesting rules from large sets of discovered association rules*. In: Proccedings of CIKM, pp.401-407.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, *11*(1), 193-197.

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, *74*(3), 036104.

Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Onnela, J. P., Kaski, K., & Kertész, J. (2004). Clustering and information in correlation based financial networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, *38*(2), 353-362.

Raeder, T. & Chawla, N. V. (2009). Modeling a store's product space as a social network. In *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in*, pp.164–169, IEEE.

Raeder, T., & Chawla, N. V. (2011). Market basket analysis with networks. *Social network analysis and mining*, *1*(2), 97-113.

Team, R. C. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.