

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO DE LA FACULTAD DE INGENIERIA

**MINERÍA DE DATOS PARA UN SISTEMA DE
RECOMENDACIONES PERSONALIZADAS**

TESIS
QUE PARA OBTENER EL GRADO DE MAESTRÍA EN INGENIERÍA DE SISTEMAS
(INVESTIGACIÓN DE OPERACIONES)

PRESENTA
DELIA LILIA LOZANO CUEVAS

DIRECTOR DE TESIS
RUBÉN TÉLLEZ SÁNCHEZ

2006



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

A mis padres Sara y Estanislao

A mis hermanas Angélica, Mariza, Grisel y Tania

A mi querido Jérôme

Índice

Introducción.....	i
Planteamiento del Problema.....	1
1.1 OBJETIVOS.....	1
1.2 HIPÓTESIS.....	2
1.3 ANTECEDENTES.....	2
1.3.1 Descripción general de la tienda.....	3
1.3.2 Consumo.....	3
1.3.3 Perfiles.....	4
1.3.4 Taxonomía de artículos.....	5
1.4 PROBLEMÁTICA.....	6
1.4.1 Comunicación con el cliente.....	6
1.4.2 Promociones.....	8
1.4.3 Análisis de información.....	10
1.5 METODOLOGÍA.....	11
1.6 ESTADO DEL ARTE.....	13
Base de Datos.....	22
2.1 CONCEPTOS BÁSICOS.....	22
2.2 ARQUITECTURA DE DATOS.....	28
2.2.1 Plataforma.....	30
2.2.2 Modelo de datos de negocio – ER.....	30
2.2.3 Estructuras principales.....	32
2.3 OBTENCIÓN DE DATOS.....	35
2.4 PREPARACIÓN DE DATOS.....	37
Diseño y Desarrollo del Modelo.....	39
3.1 MINERÍA DE DATOS.....	39
3.2 PROCESO GENERACION DE RECOMENDACIONES.....	43
3.3 ANÁLISIS DE INFORMACIÓN.....	45
3.4 ALGORITMO DE REGLAS DE ASOCIACION.....	49
3.5 CONSTRUCCIÓN DE PERFILES DE CONSUMO.....	58
3.5.1 Consumo de productos.....	58
3.5.2 Marcas.....	59
3.6 ALGORITMO DE RECOMENDACIONES.....	60
3.7 SISTEMATIZACIÓN.....	62

3.8	RESPUESTA DEL MODELO.....	65
3.8.1	Elementos de la recomendación.....	65
3.8.2	Ejemplos de listas de recomendaciones.....	66
	Aplicación.....	71
4.1	APLICACIONES.....	71
4.1.1	Enfoque natural.....	72
4.1.2	Enfoque inverso.....	73
4.2	MEDIOS DE CONTACTO.....	74
4.2.1	Punto de venta.....	74
4.2.2	Telemarketing.....	75
4.2.3	Correo tradicional.....	75
4.2.4	Correo electrónico.....	75
4.2.5	Página en Internet.....	76
4.3	APLICACIONES COMPLEMENTARIAS.....	76
4.4	PRUEBAS.....	77
	Conclusiones.....	80
	Anexo A.....	82
	Anexo B.....	90
	Bibliografía.....	105

Introducción

En el mundo actual, donde existe un mayor y mejor acceso a la información, donde los cambios ocurren de forma rápida y decisiva, los hábitos, necesidades y expectativas del consumidor cambian día con día, la innovación permanente de productos y el “marketing” juegan un papel fundamental para las empresas, haciendo que la competencia por satisfacer las necesidades del cliente resulte una prioridad.

Uno de los sectores más desafiantes es el de las tiendas departamentales, ya que al contar con productos, promociones y precios iguales, la competencia se torna cada vez más inexorable, haciendo que la diferencia se encuentre no sólo en el precio sino en los beneficios adicionales o el servicio.

En respuesta a ello, las empresas han llevado a cabo propaganda dirigida, sin embargo, el gasto y el esfuerzo pueden resultar ineficiente y generar irritación o cansancio sino se hace de manera planeada. El “marketing” dirigido logra este acercamiento, que puede ser tan o más efectivo que los medios masivos. Existen varias opciones como la publicidad en el punto de venta, telemercadeo, envíos, correo electrónico o materiales para la fuerza de venta, pero esto debe manejarse con cautela, ya que se puede incomodar al cliente sino es de su interés, por ejemplo, la publicidad de ofertas de juguetes infantiles para un hombre soltero no resulta tan atractiva como para un padre de familia.

El contar con herramientas que ayuden a entender al cliente, a ofrecerle productos o servicios acorde a sus necesidades, gustos o hábitos y a establecer un contacto más cercano, es necesario para establecer lazos que refuercen la lealtad y que conviertan la relación con el cliente más valiosa y duradera.

En respuesta a ello, se pensó en generar información de interés particular para cada cliente específico, con el fin de sentar las bases en el fortalecimiento de vínculos, teniendo la posibilidad de ofrecerle lo que desea o le gusta, en el tiempo y forma correcta. Esta información es generada mediante el uso de herramientas computacionales, una Base de Datos como plataforma para el almacenamiento y manejo de información, y una herramienta de Minería de Datos para el análisis de asociaciones entre productos, usando el algoritmo de Asociaciones mejor conocido como “Market Basket Analysis”.

Dada la importancia de conocer a fondo el entorno de la información y del estilo de consumo de los clientes, en el capítulo I se da un panorama general del problema y se explican los detalles tanto del funcionamiento de la tienda como de los canales de comunicación con el cliente y la distribución de productos, así como la relación tienda-cliente, también se revisan casos similares en la literatura con la finalidad de comprender y analizar de forma óptima el problema.

En el capítulo II se describe la plataforma de Base de Datos, desde el modelado de información hasta la preparación de los datos, tomando como fuente el consumo detallado de cada cliente en el tiempo.

En el capítulo III se diseña y desarrolla el proceso que permite conocer los productos recomendables para cada individuo. Con la ayuda de un modelo de Minería de Datos se determinan las relaciones que guardan los productos entre sí, cabe mencionar que la Minería de Datos juega el papel más importante, ya que representa la base para conocer los vínculos entre artículos o mercancías y más importante aún, conocer el tipo de vínculos, qué tan fuertes o débiles son y la conglomeración de los mismos. Posteriormente se completa el proceso para generación de recomendaciones personalizadas, destacando las asociaciones existentes, la construcción de perfiles de consumo y del tipo de relación entre productos.

De forma específica, se busca, a través del modelo sistematizado, generar listas de recomendaciones personalizadas para cada cliente, de acuerdo a su perfil de consumo, las afinidades de la cartera de productos ofrecida, el tipo de relación entre productos y la matriz de marcas de moda y precio.

Aunado a lo anterior, es primordial, además de conocer el producto adecuado para cada cliente, establecer el mejor canal de comunicación para proporcionar dicha información al cliente, por ello, en el capítulo IV se analiza diversos medios de contacto y la conveniencia de cada uno de éstos, con el fin de lograr un contacto de mayor precisión y calidad con los clientes.

Un buen servicio a un cliente puede ser un elemento muy poderoso, los mayores esfuerzos se ponen en identificar y satisfacer las necesidades del consumidor, es importante que el cliente reciba sugerencias de productos que le resulten atractivos y no de aquellos que están fuera de sus expectativas o gustos. Si al cliente se le recomienda un producto atractivo, el cliente se sentirá entendido y conocido, y es probable que lo conduzca a tener una liga más fuerte con la tienda, generando así lo más valioso para cualquier empresa, fidelidad del cliente.

Capítulo I

Planteamiento del Problema

INTRODUCCIÓN

Los consumidores de nuestros días han cambiado sus hábitos de compra y comportamiento, cada día la competencia es más intensa y es dirigida en diferentes sentidos, uno de estos casos son las tiendas departamentales. La imagen es importante y muchas veces es utilizada para fortalecer la fidelidad y confianza de los clientes y garantizar así su supervivencia largo plazo.

La tienda departamental de la cual trata este trabajo tiene una imagen importante en la sociedad mexicana, ha existido desde hace más de un siglo y tiene mucho prestigio.

Por otra parte se tiene el cambio en los hábitos de consumo de las personas, esto hace que cada vez sea más complicado tener clientes fieles, ya no existen más los clientes cautivos, y con la competencia cada vez más dura resulta complejo hacer que una promoción sea atractiva; cada vez el cliente es más conocedor y globalizado, por tanto más exigente. Esto hace que la comunicación con el cliente tienda a ser más personalizada, más directa, con el fin de que el cliente sea más receptivo, se interese y responda a la publicidad.

Este capítulo explica el problema de estudio, así como el contexto en el que está envuelto. El capítulo está dividido en 6 partes, en la primera se explica el objetivo del trabajo y en la segunda, la hipótesis. La tercera parte explica el contexto general de la tienda, se describe cómo es el consumo de sus clientes, así como los perfiles de los consumidores, por otra parte se expone la clasificación de los artículos que ofrece la tienda. En la cuarta parte se describe la comunicación con el cliente, el tipo de promociones que se ofrecen en la tienda, así como el manejo y almacenamiento de la información. La quinta parte explica la metodología utilizada para la generación de las recomendaciones personalizadas y el estado del arte.

1.1 OBJETIVOS

El objetivo de este trabajo es generar listas de recomendaciones individuales de productos para los clientes de una tienda departamental, apoyándose en los hábitos de consumo individuales, las afinidades de la cartera de productos ofrecida, y la matriz de marcas de moda y valor (que ayudan a posicionar una marca por nivel de precio y tipo de moda), con el apoyo de algoritmos de minería de datos, con el fin de conocer las relaciones entre productos.

Se desea hacer recomendaciones personalizadas a los clientes de una tienda departamental, de los productos que en ella se ofrecen, de tal manera que las recomendaciones sean afines al perfil de consumo de cada cliente, de esta manera se logra ofrecer productos clientes que tienen

correspondencia más cercana con éstos, es decir se tiene un contacto de mayor precisión con los clientes y los productos a ofrecerles.

Para lograr este contacto es necesario analizar más a fondo el entorno de información, tanto del negocio mismo como del consumo de los clientes. Esta análisis es realizado mediante el uso de una herramienta una base de datos y el algoritmo de asociaciones (Market Basket Análisis, o Análisis de Cesta de Compra) de minería de datos, tomando como fuente los hábitos de consumo de cada cliente, y la relación que guardan los productos entre sí, con el fin de alimentar un modelo que permita recomendar productos afines a los consumidores. Cabe mencionar que la minería de datos juega el papel más importante, ya que es la base para conocer los vínculos entre artículos o mercancías y más importante aún conocer el tipo de vínculos, que tan fuertes o débiles son y la conglomeración de los mismos.

Se espera que se opere en el piso de venta, y cuando el cliente acuda a comprar algún producto y pague con su tarjeta de crédito, al momento de deslizar la tarjeta directamente se consulte en una base de datos la mejor recomendación de ese cliente, en ese momento mostrar en la terminal punto de venta un mensaje que el(la) vendedor(a) podrá dar a conocer al cliente como una recomendación exclusiva y única para esa persona.

El objetivo final es hacer sentir al cliente una motivación para comprar algo que le resulte interesante o que sea de su gusto, que le haga “clic”, que se sienta como un cliente entendido y conocido, para que lo conduzca a tener una liga más fuerte con la tienda, generando así fidelidad.

1.2 HIPÓTESIS

La utilización de la Minería de Datos en el conocimiento de las relaciones entre productos para generar recomendaciones personalizadas, y el ofrecimiento de artículos específicos a un cliente, que sean de su gusto o preferencia, con base a los patrones encontrados a través de un algoritmo de Minería de Datos, incrementa la probabilidad de que el cliente realice una compra, comparado con el hecho de sólo ofrecer productos de manera general sin conocer la relación puntual entre ellos y el vínculo con cada cliente.

1.3 ANTECEDENTES

Las tiendas departamentales son muy comunes y populares hoy en día, en ellas se puede encontrar todo tipo de cosas desde galletas, pasando por ropa y joyas hasta cocinas integrales o motocicletas, esta mezcla de productos las hace muy interesantes y atractivas, esto permite a un cliente la posibilidad de comprar un abanico de productos diversos en el mismo espacio.

En México hay tiendas departamentales de todos los niveles y dirigidas a diversos públicos. Este trabajo se desarrolla en una tienda departamental cuya imagen es la exclusividad, lo cual resulta muy atractivo tanto para las personas de alto nivel como para las de medio, y especialmente este

último resulta ser el tipo de cliente más atraído, a continuación se describe con detalle el funcionamiento de la tienda, así como el consumo de sus clientes.

1.3.1 Descripción general de la tienda.

Este trabajo se desarrolla en una cadena de tiendas departamentales, el concepto es el de una tienda de lujo y es vista frecuentemente como una tienda para conseguir un estatus alto dentro de la sociedad, en ella se ofrecen tanto productos como servicios de diversos tipos, tales como: ropa, zapatos, accesorios, joyería, perfumería, muebles, productos de hogar, línea blanca, electrónica, cómputo, libros, discos, juguetería, comida gourmet, también tiene restaurantes, diferentes boutiques de ropa, agencia de viajes, telefonía, servicios de diversos tipos y módulos de seguros.

Debido a la naturaleza y el concepto de la misma, una de las cosas más importantes es brindar una buena atención y un excelente servicio, de tal manera que el cliente se sienta plenamente satisfecho con su compra. Una de las partes más importantes es que está enfocada a los productos de última moda, pero también tiene los básicos.

Generalmente los clientes son atraídos por la excelente publicidad buscando artículos de moda, básicamente enfocados a la ropa, pero en muchos casos desconocen la gama de productos y servicios que se ofrecen.

La tienda se ve de manera pública como muy exclusiva y glamorosa, dirigida principalmente a la mujer. Vende productos de alta calidad, de marcas muy prestigiadas a nivel mundial, y también productos de niveles de precios accesibles a la clase media, que son los productos de marcas propias o exclusivas.

1.3.2 Consumo

El consumo dentro de la cadena es un tema que se ha analizado desde diversos ángulos, y gracias a que la tienda tiene su propia tarjeta de crédito se cuentan con datos de los clientes, de tipo crediticio, demográficos y de consumo. Los puntos más importantes y relacionados al presente trabajo son los siguientes:

- Frecuencia de compra.

En general los clientes no acuden con tanta frecuencia, ya que los artículos que se venden no son de primera necesidad como los de un supermercado, por lo cual las visitas quincenales o mensuales son escasas. Pero existen productos de gourmet o perfumería que requieren resurtido en períodos de tiempo más corto que el resto, como pueden ser zapatos, ropa o equipos electrónicos.

- Género de los clientes.

A pesar de que la publicidad está prácticamente dirigida a mujeres, también se tiene una cantidad considerable de clientes hombres (cerca del 50%), aunque un poco más de la mitad de los clientes son mujeres, los hombres tienen montos más altos de gasto que las mujeres, lo cual hace que el gasto global de cada caso sea prácticamente mitad.

- Utilización de Crédito.

Se mide la utilización de la línea de crédito de cada cliente en periodos de 12 meses, y se ha identificado que la mayoría de los clientes subutilizan su línea de crédito, es decir que gastan un porcentaje bajo de su capacidad de compra.

- Tipos de consumo.

- Compras de alto valor.

Algunos clientes planean compras de montos altos no periódicas, este caso se refiere a aquellos clientes que compran productos de alto valor, tal como un electrodoméstico, una computadora, un artículo de electrónica o un incluso un viaje. En este caso, normalmente el cliente realiza la compra para pagar de manera diferida en varios meses, y no compra nada más por un período de tiempo largo o medio.

- Consumo general.

Los clientes consumen artículos de vestir o de moda, de valor medio o bajo, muchas veces planean sus compras y las realizan apegándose a su plan original. Esto puede generarse debido al desconocimiento de la gama de productos que ofrece la tienda.

- Consumo por impulso.

La distribución de cada tienda está diseñada de tal manera que los departamentos más atractivos son los que se encuentran en las zonas de tránsito intenso, mientras que el resto se ubica en lugares menos visibles, esto se debe a que muchas veces se compra por impulso, tal vez un cliente pensaba comprar algún producto específico, pero en el momento que entra a la tienda o caminando por las zonas céntricas, recuerda que necesita o siente el deseo de comprar un artículo que está viendo, aunque no lo necesite o no haya planeado comprarlo, en ese momento adquiere el producto.

1.3.3 Perfiles

Aunque los vendedores dentro de la cotidianidad hacen recomendaciones a los clientes, estas son de manera general, y muchas veces son basadas en ofertas del momento o en productos nuevos, pero sin tomar en cuenta el perfil o preferencias de cada persona.

Es importante hacer notar que no sólo existe comunicación directa con el cliente a través de los vendedores, también se hacen propaganda por otros medios, como son el envío de invitaciones,

folletos o catálogos por correo, mensajes y mailers en el estado de cuenta, telemarketing o e-mailing. Sin olvidar los medios masivos como son anuncios espectaculares, periódicos, radio, tv, pero estos últimos resultan muy costosos.

Es importante recordar que cada persona es diferente, existen muchas maneras diferentes de pensar, niveles de ingresos, estilos de vida y gustos, por lo cual es imposible recomendar lo mismo a todas las personas y que resulte igual de atractivo o surja un efecto similar.

Generalmente las personas tienen un estilo, por ejemplo si se trata de una persona adulta y madura, que compran ropa de estilo clásico, sus marcas favoritas son clásicas, escucha música clásica, seguramente su casa está decorada de manera clásica y su auto también es clásico. De manera similar ocurre con una persona joven y relajada, que compra las marcas más vanguardistas, la música de última moda, tenis novedosos, video juegos, seguramente su estilo de vida es más vanguardista o contemporáneo.

Esto lleva directamente a pensar que ofrecer una bicicleta de montaña a un anciano, ofrecer vino a un abstemio, una computadora a una persona conservadora en tecnología, es enfocar recursos y esfuerzos a algo de lo que se tendrá un resultado muy bajo o nulo.

Derivado de esto, se pensó en estudiar el perfil de compra de cada persona, de tal manera que se detecte su historial de consumo para conocer su manera de comprar, y por consecuencia su estilo o perfil de consumo, para hacer recomendaciones congruentes con su naturaleza, mismas que pueden ser tan generales o detalladas como se desee, hasta llegar al punto de efectuar sugerencias que sean únicas para cada persona y que resulten atractivas para todos y cada uno de los clientes.

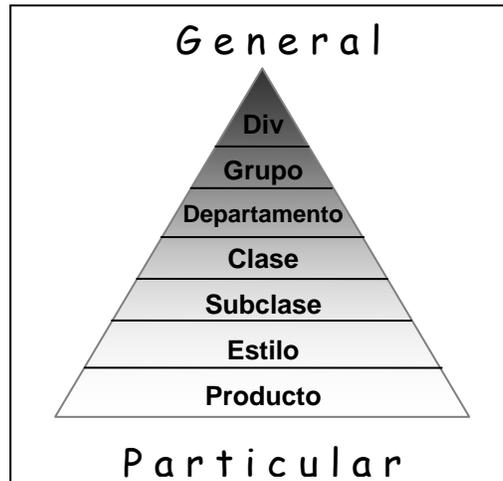
Por ejemplo si una persona compra vino, se le podría recomendar un queso que corresponda con el vino, o si compra una camisa se le podría recomendar una corbata de su marca favorita, no necesariamente en el momento en el que está comprando el primer producto, puede ser en la siguiente visita a la tienda o incluso con una comunicación vía correo, e-mail o telefónica que derive en una visita a la tienda ya sea física o virtual.

1.3.4 Taxonomía de artículos

Dentro de la tienda, los productos se clasifican para el manejo administrativo, lo cual resulta muy práctico para los análisis de efectividad dentro de la organización, pero desde el punto de vista del cliente no es práctico, por lo cual se plantea una modificación.

A continuación se muestra la taxonomía de los productos en la figura 1.1, es una estructura piramidal que tiene los productos en la base y conforme se agrupan se sube dentro de la pirámide, hasta llegar a división. El nivel más importante, para este caso es el departamento, ya que casi en todos los casos está dividido en tres partes de acuerdo al manejo administrativo, por ejemplo el departamento Damas tipo1, Damas tipo 2 y Damas tipo 3, pero desde la perspectiva de un

consumidor es sólo Damas, por lo cual se propone agregar un nivel a la pirámide agrupando este tipo de clasificación.



Taxonomía de productos

Figura 1.1

Después de conocer de manera general los puntos más importantes la tienda que influyen en este trabajo, en la siguiente sección se habla de la problemática actual dentro del negocio.

1.4 PROBLEMÁTICA

Lo ideal para cualquier negocio es conocer a sus clientes de manera profunda, y con esto entender sus necesidades, gustos y posibilidades de compra, pero lo más valioso es hacer uso de esta información para comunicarse con el cliente, ofrecerle los productos y/o servicios que espera.

Además de conocer al cliente, sus gustos, necesidades e historial de compra (almacenados en una base de datos), en este proceso se involucran varios elementos extras como son la factibilidad de comunicación, la información que se envía al mismo (los productos y/o servicios que se le ofrecen) y la capacidad de medir el efecto que causa.

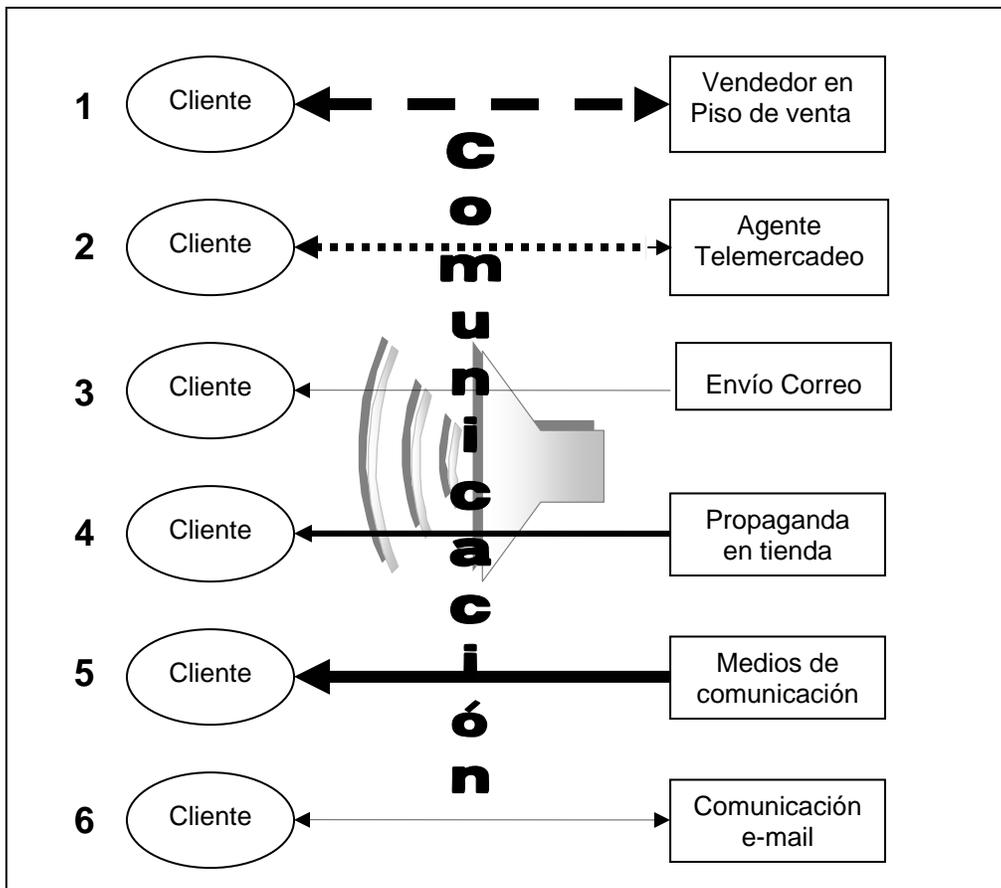
1.4.1 Comunicación con el cliente

Actualmente debido a la naturaleza y a las dimensiones del negocio se desconoce al cliente, es decir que se tiene una relación impersonal con éste. Los clientes son tratados de manera colectiva y en caso de hacerse una recomendación, ésta es muy general, ya que se desconocen totalmente los gustos, deseos y compras previas de cada cliente específico. Los tipos de comunicación que existen están clasificados en seis tipos de acuerdo a la naturaleza de los mismos (Figura 1.2):

1. Piso de venta.

La comunicación más efectiva es en el piso de venta, ya que es una comunicación directa con el cliente, además se tiene la enorme ventaja de que el cliente se encuentra en la tienda, razón por la cual este tipo de comunicación resulta el más importante, directo y con mayor posibilidad de éxito que el resto.

En algunos casos cuando un(a) vendedor(a) atiende continuamente a un cliente llega a conocer sus gustos y preferencias, pero estos casos son reducidos, además de que la rotación de personal y horarios son frecuentes.



Tipos de comunicación

Figura 1.2

2. Telemarketing.

La comunicación telefónica es totalmente indistinta, no existe la posibilidad de conocer al cliente de manera individual como en el caso del piso de venta.

El telemarketing resulta ser la segunda comunicación más efectiva, porque se puede tener comunicación verbal directa con el cliente, esto podría retroalimentar al agente telefónico, pero el proceso está diseñado sólo para conocer si el cliente recibió o no el mensaje.

3. Envío correo.

El correo directo no es un tipo de comunicación efectiva, ya que no existe retroalimentación del cliente, es totalmente nula. Usualmente se hacen envíos masivos de propaganda e invitaciones, de manera masiva y no existe posibilidad de conocer si el cliente está o no interesado o incluso en algunos casos si recibió el correo. También existe el problema del exceso de “propaganda basura”, lo que ha provocado que no se tenga interés de abrir el correo con propaganda y frecuentemente termina en la basura sin ser siquiera visto.

4. Propaganda en tienda.

En las tiendas se colocan letreros donde se anuncian ofertas, promociones y eventos, los clientes estando dentro de la tienda se enteran, tiene la ventaja de que el cliente ya se encuentra dentro de la tienda, pero no hay manera de conocer si resulta o no atractivo.

5. Medios de comunicación masivos.

La comunicación a través de los medios de comunicación masivos, como son televisión, radio, periódicos y espectaculares. Este caso es uno de los más eficientes pero al mismo tiempo muy caro y no existe ninguna posibilidad de medir el impacto.

Aunque es importante señalar que la campaña publicitaria de la tienda es muy exitosa y ha logrado la retención de la marca de manera muy efectiva, pero es totalmente impersonal.

6. Comunicación vía e-mail.

Este tipo de comunicación resulta poco eficiente (por ahora), se podría tener retroalimentación pero el proceso actual no lo contempla, por otra parte, resulta parecido al envío por correo, a todos los clientes se les envía se envía lo mismo, igualmente por el exceso de “spam” muchas veces es desechado sin siquiera ser visto.

1.4.2 Promociones

Como ya se ha mencionado, los productos y/o servicios que se ofrecen al cliente, es decir las recomendaciones, se hacen con base en las promociones, descuentos y eventos que la tienda realiza periódicamente, no son pensadas con base en las necesidades y gustos de cada cliente de manera particular sino de manera general.

Usualmente se basan en la experiencia de varios años, si una promoción ha sido exitosa en diferentes períodos se continua, sino se elimina. Y las nuevas surgen por ideas generales de tendencias o necesidades.

Existen diferentes tipos de promociones y eventos, los cuales se dan a conocer a los clientes con el fin de que se adquieran productos y/o servicios relacionados.

En general se pueden clasificar en cinco tipos diferentes:

1. De temporadas

Algunas promociones están basadas en estacionalidad y moda, al fin de las temporadas primavera-verano y otoño-invierno se hacen descuentos masivos, con el fin de apresurar la venta de productos para dar paso a las mercancías de la nueva temporada.

Otro caso es el inicio de temporada, se hacen eventos para presentar las nuevas tendencias de la moda, y con esto estimular la actualización y compra de productos nuevos.

2. Festividades

Se refiere a todo aquello que tiene que ver con festejos generalizados, por ejemplo el día de la madre, el día del padre o semana santa, cercano a estas fechas se ofrecen descuentos y/o productos específicos para esas conmemoraciones.

3. Días no laborables

Este caso es sobre días festivos donde no se labora, la idea es aprovechar que las personas no deben trabajar y tienen tiempo libre para ir de compras, generalmente se hacen descuentos generales en estos días.

4. Territoriales

Se refiere a casos de una tienda específica, esto se hace por el aniversario de la tienda, o por un festejo local del área donde está ubicada. Esta dirigida a los clientes que geográficamente son cercanos a la tienda.

5. Productos especializados

Otro caso son las promociones de un conjunto de productos o servicios que tienen un denominador común como puede ser el tipo, el departamento o la marca, aunque también puede manejarse en el contexto de un solo producto, esta clase de promociones es la más especializada y es donde se deben enfocar más los esfuerzos de la comunicación con el cliente.

Otro caso que es el de las promociones para clientes pertenecientes a un club, resulta más simple y está totalmente dirigido y pensado para que sea atractivo para los miembros, pero existe un área de oportunidad en la invitación para formar parte del club a nuevos clientes.

Los tipos de productos y/o servicios que se le ofrecen al cliente no están relacionados con la individualidad de cada uno, así como tampoco se le da seguimiento a la efectividad de la comunicación son los mismos y lo más importante, el efecto que causa.

Este proceso requiere de una infraestructura robusta, apoyada en la tecnología, en la capacidad de almacenamiento de grandes volúmenes de información, y la eficiencia en la consulta, operación y

manejo de la misma. A continuación se habla sobre este punto, base fundamental en el desarrollo del proceso.

1.4.3 Análisis de información

La tienda cuenta con una tarjeta de crédito, y un área enfocada al manejo, administración y mantenimiento de la información derivada de ésta, se tienen bases de datos con información de los clientes, así como las transacciones o compras realizadas, e información crediticia del uso de la tarjeta. Por otra parte, existe un área que maneja toda la parte que tiene que ver con los productos, desde la planeación, la adquisición, el envío y distribución en tiendas, inventarios y las transacciones de todos los clientes, tarjeta-habientes y no tarjeta-habientes. También se tienen bases de datos que permiten el almacenamiento, manejo y administración de esta información, esto permite conocer el detalle de las compras a nivel de cada producto, con particularidades tan específicas como color y talla.

En ambos sentidos, la información se concentra en dos depósitos de datos diferentes, que son bases de datos donde se almacena una colección de información orientada a un tema, llamados típicamente Data Warehouse que es como de ahora en adelante se hace referencia, que no tienen ningún tipo de conexión entre ellos, estas dos bases de datos son muy grandes y contienen millones de datos, están enfocadas al manejo de información en dos sentidos diferentes, pero no se integran. Como ya se mencionó, con las bases de datos existentes se puede conocer información en dos sentidos: 1) Los datos demográficos y crediticios del tarjeta-habiente y sus montos de compra, y 2) El detalle de compra realizado por una tarjeta de crédito, pero sin conocer los datos del cliente.

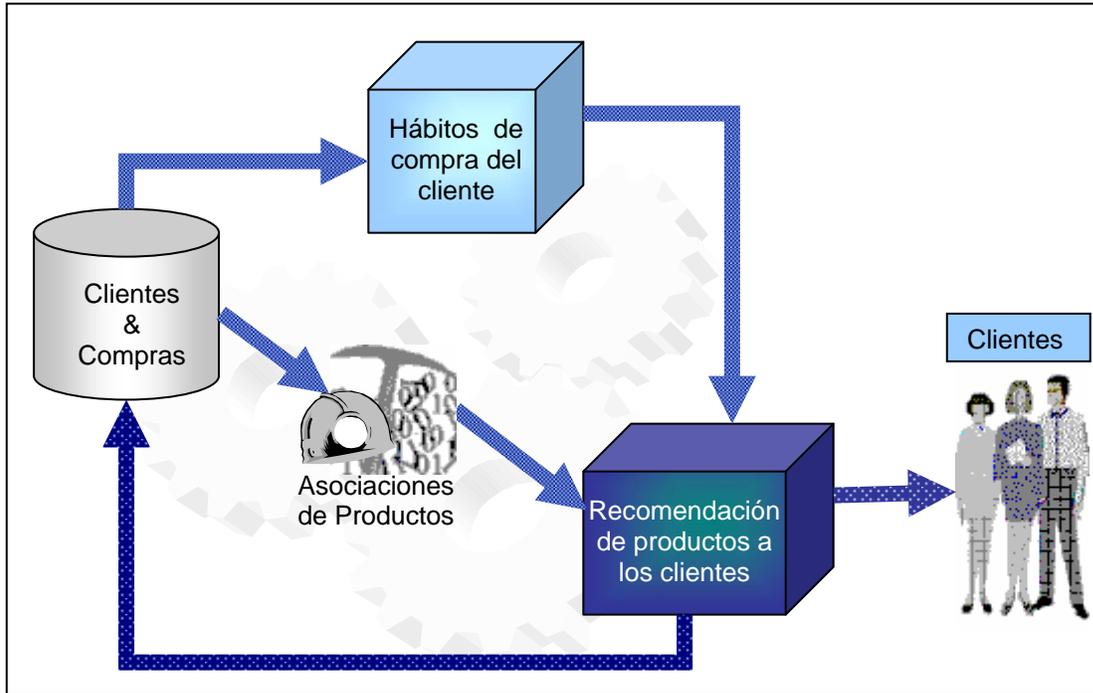
Para conocer el detalle de compra de un cliente y los datos de la persona es necesaria la integración de información proveniente de las dos vertientes, por lo cual se requiere la definición de un modelo de datos nuevo, es decir la creación de una base de datos que integre la información del cliente con su detalle de compra, que resulte eficiente y rápida la obtención de datos en términos computacionales. Normalmente un modelo de datos define: fuentes, tipos, contenido, descripción y usos, pero esto se revisa en el capítulo 2.

Para la mayoría de las aplicaciones de minería, el modelo de datos requerido toma la forma de un conjunto de archivos o tablas, con un registro para cada sujeto de la investigación con la información a estudiar, en este caso para cada cliente y conjunto de productos. Cada registro incluye una o muchas variables, que pudieron derivarse de fuentes múltiples.

Poseer un modelo de datos significa contar con la estructura necesaria, que permitirá conocer la información ligada a cada cliente en todos los sentidos, y la obtención de datos implica seguir las etapas de identificación, colección, filtrado y agregación de datos, de acuerdo al diseño requerido por el modelo de datos y las técnicas de análisis que pueden ser utilizadas. De esto se habla con mayor profundidad en los capítulos 2 y 3.

1.5 METODOLOGÍA

El proceso seguido en la generación de recomendaciones personalizadas es descrito a continuación (Figura 1.3).



Metodología para la generación de recomendaciones

Figura 1.3

1. Fuente de información que alimenta el proceso.

En primer lugar, es necesario contar con una fuente de información que genere de manera simple y rápida la estructura de datos que es utilizada para alimentar el proceso de generación de hábitos de compra y el de las recomendaciones.

Actualmente se tienen dos fuentes de información, pero están totalmente separadas y como fueron creadas para propósitos diferentes, resulta muy complicada la explotación de información dentro de las mismas, además, sino se crea una estructura de datos dentro de alguna de ellas en la que se agreguen los datos complementarios, es imposible generar la información, por lo cual debe consolidarse en una sola fuente, que además debe ser de fácil explotación. Este es un proceso largo y complicado debido a los requerimientos técnicos, y es explicado a detalle en el capítulo 2.

2. Proceso para conocer los hábitos de compra de cada cliente.

Debe ser basado en el historial de compras de cada cliente, no se trata de hacer una lista de productos que el cliente compró durante un período, si no de agruparlos y hacer que con esta agrupación se muestren los hábitos de compra de cada uno. El proceso se desarrolla en el capítulo 3.

3. Proceso para la generación de recomendaciones personalizadas.

Requiere tomar la experiencia de compra de otros clientes para encontrar la relación existente entre productos, y tomando por otra parte los hábitos de consumo de los clientes, hacer una integración de ambos datos, partiendo de los artículos comprados por el cliente, verificar cuáles son los que están más relacionados con ese conjunto de artículos y con base en esto hacer una recomendación que sea única para cada cliente, recomendando sólo aquellos productos que no han sido adquiridos por el cliente, así como utilizar la marca favorita del cliente y la matriz de marcas para recomendar la marca favorita y en caso de no existir un producto de esta marca recomendar una marca que sea equivalente tanto en el tipo de moda como en el nivel de precio. Tomar como base una calificación que permita identificar los productos mas recomendados de los menos recomendados, con base en las relaciones descubiertas por la Minería de Datos, con el fin de que existan niveles de recomendación de productos.

Elección de la técnica de minería de datos es el punto más importante y en el cual estará basado toda la fortaleza del proceso, así como la forma en la que debe ser aplicada.

La empresa cuenta con un software de Minería de Datos llamado “Intelligent Miner for Data”, que constituye la herramienta principal en el desarrollo de este trabajo, los algoritmos contenidos en el software son los siguientes: clasificación, reglas de asociación, segmentación, predicción, patrones secuenciales y secuencias similares. El software también incluye procesos para la detección de diferentes tipos de correlaciones o patrones dentro de la base de datos: análisis de componentes principales, análisis de factores y regresión.

El reto no es sólo determinar cuál es la técnica idónea a utilizar, también el cómo debe aplicarse, pero de esto se habla más a fondo en el capítulo 3.

4. Almacén de resultados.

Es imprescindible tener un almacén de datos, donde se puedan guardar las recomendaciones, así como para indicar cuando ya se ha hecho cada recomendación a cada cliente, con dos fines:

- 1) Control de las recomendaciones.

En dos sentidos, tanto de productos como de los tiempos y la frecuencias.

- 2) Medir el efecto de las recomendaciones.

Una vez que la recomendación haya sido dada a conocer al cliente, es conveniente medir el efecto de la misma, aquí se tiene que pensar muy bien en el período en que se puede medir el

efecto, aún cuando la recomendación la haga el vendedor en el piso de venta, no necesariamente el cliente va a responder en ese momento, puede ser que prefiera esperar al fin de semana o a la quincena para realizar la compra. Por otra parte, si la recomendación se hace por un medio indirecto, también debe esperarse un periodo de tiempo razonable para verificar el efecto de la misma, esto se ve con más detalle en el capítulo 2.

Los procesos mencionados anteriormente deben desarrollarse bajo la premisa de que cumplan con las siguientes características:

- Capacidad de manejo de altos volúmenes de información.
- Procesos automatizados, de tal manera que el cálculo de los mismos no tenga procesos manuales y se generen de manera mecánica al cumplirse requerimientos de tiempo o de información.
- No debe existir duplicidad de información, es decir que la base de datos debe ser única y congruente con la información existente dentro de las otras bases de datos.
- Capacidad de evaluación, el resultado del proceso debe ser medible, con el propósito de hacer ajustes o mejoras en el futuro, y para conocer el efecto en cada cliente, así como utilizar otros tipos de datos y que sea útil en futuros análisis.
- Respuesta rápida, durante la generación de información como durante el uso de la misma, por ejemplo en el piso de venta, el cliente sólo va a esperar el tiempo requerido para que la transacción sea autorizada al deslizar su tarjeta, ese tiempo debe ser suficiente para mostrar la recomendación. Al deslizar la tarjeta, el tiempo en que viaja el número para la autorización del crédito, debe conjuntamente viajar la información de la recomendación y retornar simultáneamente con la autorización y con la recomendación, para que ésta aparezca en la pantalla del vendedor y la dé a conocer al cliente.
- Periodicidad, las recomendaciones deben ser generadas de manera constante, con el fin de que se actualicen continuamente para que sean vigentes todo el tiempo.

1.6 ESTADO DEL ARTE

Un sistema de filtro colaborativo, como Ringo (Shardanand, U. & Maes, P., 1995), Phoaks (Terveen L. G. et al, 1997), GroupLens (Konstan et al., 1997), MRS (Chen H.C. & Chen A.L.P., 2001), Sitieseer (Rucker J. & Polanco M. J., 1997), NewsWeeder (Lang K., 1995), (Allen, R. B., 1990), (Hill, W. C. et al, 1995), funcionan recabando explícitamente calificaciones que los usuarios dan a los artículos en cuestión. Para calcular las recomendaciones “con la palabra en la boca”, se ha sugerido el uso de segmentaciones de perfiles de usuarios con el fin de definir modelos de

perfiles (Ungar & Foster, 1998). Esperando que todos los usuarios hagan lo mismo y tengan el mismo beneficio.

Las reglas de asociación minería de datos han sido estudiadas con el algoritmo A priori (Agrawal R. & Srikant R., 1994), (Srikant R. & Agrawal R., 1995) analizan la relación entre productos con un gran número de transacciones utilizando la taxonomía de los mismos, encuentran asociaciones entre artículos en todos los niveles de la taxonomía, y presentan una métrica que usa la información dentro de la taxonomía, de tal manera que elimina las reglas redundantes.

(Breese et al., 1998) dicen que las técnicas de segmentación generalmente producen recomendaciones menos personalizadas, o con peor precisión que algoritmos basados CF.

(Good N, et al, 1999) usan filtrado de información IF y colaborativo CF para mejora la selección de artículos, IF se enfoca en los intereses de un perfil de usuario, es valioso cuando no hay información previa; mientras que CF identifica gustos de usuarios similares, pero no considera la satisfacción, se encuentran los gustos generales. Cada técnica tiene ventajas y limitaciones pero combinarla puede ser benéfico. La técnicas de retroalimentación de información (IR) se identifica con gustos efímeros, como buscar un tema en una librería (Belkin & Croft 1992).

(Balabanovic & Shoham 1997) combinan IF y CF en 3 sentidos: usan filtrados inteligentes aprendidos de recomendadores humanos, se aplica a comunidades pequeñas y se evalúa el uso simultáneo de varios filtros.

Muestran como dentro de un marco CF se puede combinar IF y las opiniones de los usuarios para producir mejores recomendaciones. Muestran que CF puede usarse combinando agentes humanos y métodos. Explotan el valor de IF y CF para crear un sistema efectivo de recomendaciones de películas de cine.

(Friedman, B., 1999) dice que los individuos podrían ser informados sobre problemas y beneficios de participar en una actividad particular, y podrían decidir sin participar o no.

Los sistemas de recomendaciones son procesos automatizados de experiencias de una comunidad, dadas a conocer a un usuario que va a tomar una decisión. El sistema sistemas determina los datos a recabar, su almacenamiento, su acceso, su retroalimentación, así como cuando y como presentarlos al usuario. Los sistemas pueden beneficiar al usuario cuando se emplean correctamente, sin embargo, la confianza del usuario en interacciones web puede no ser alta.

(Lueg, C., 1999) Discute tópicos sobre el filtrado colaborativo, trata de entender la naturaleza de las recomendaciones, de acuerdo a las recomendaciones despersonalizadas de las compañías y el éxito de las recomendaciones anónimas, como es el caso de Amazon. En cuanto al filtrado colaborativo y los grupos, se requiere mayor peso para deducir usuarios proveedores, se tiene el problema del inicio frío.

(Baudisch P., 1999) Trata un camino diferente introduciendo descripciones de contexto dentro de sistemas colaborativos. Propone unir las recomendaciones no en el sentido de combinarlas, si no

en el sentido relacional, para integrar sistemas colaborativos con sistemas de información de contexto y generar nueva funcionalidad, disponible directamente para los usuarios.

(Wolf et al. 1999) usa una técnica con graficas de nodos y ramas que indican la similitud entre usuarios, las predicciones se hacen al desplazarse por el árbol y combinando la opinión de usuarios afines, con esta técnica se produce una mejor predicción con CF.

Las reglas de asociación para recomendaciones de espacios también han sido estudiadas (Ramakrishnan N. & Ribbens C., 2000), identifican puntos importantes en la minería de espacios con un algoritmo de programación dinámica.

Más adelante se ha estudiado más respecto a la correlación artículo-artículo (Sarwar et. Al, 2001). (Adomavicius G. 2001) argumenta que la aproximación de contexto provee sólo entendimiento y está limitado sólo al comportamiento, es necesario capturar líneas de comportamientos basados en su historia transaccional. El sistema recomendador contextual está limitado y no tiene la capacidad de decidir si se le hace o no una recomendación adicional al usuario.

La literatura en sistemas recomendadores distingue entre la aproximación colaborativa y contextual y considera la combinación como una mejor aproximación, que puede basarse en (1) entender el comportamiento del consumidor, y (2) en preferencias de otros consumidores similares. Es importante desarrollar métodos que combinen el perfil de comportamiento y las aproximaciones colaborativas en in método integral.

(Tan P-N. & Kumar V., 2001) aplican las reglas de asociación a datos Web flujo-clic. La idea es encontrar pares de páginas que son negativamente asociadas entre ellas, pero son positivamente asociadas con otro conjunto de páginas llamadas el mediador. Estos pares de páginas se llaman para ser indirectamente asociativas vía su común mediador.

(Adomavicius G. 2001) presenta una propuesta de un sistema recomendador en tiempo real con un DW, para trabajar en conjuntos multimedia, soportados perfiles con las capacidades OLAP, para hacer recomendaciones multidimensionales, que incluyen dimensiones múltiples, así como perfiles extensivos y agregaciones.

(Lawrence R.D. et al, 2001) describe un sistema de recomendaciones diseñado para sugerir nuevos productos a clientes de supermercado basado en compras previas. A través de una red de dispositivos de servicios móviles PDAs, conectado al sistema SmartPad.

Dos paradigmas principales para el filtrado emergieron: filtrado contextual y colaborativo.

Filtrado contextual: Típicamente, artículos (items) y perfiles se representan como vectores en el espacio de rasgos y sus similitudes son calculadas con una medida estándar de distancia, como un coeficiente. Esta aproximación tiene su raíz en el modelo base-vector de retorno de información (IRetrieval).

El sistema propuesto usa dos fuentes de información. Primero, la asociación de minería para datos de compras de los clientes para derivar relaciones de clases y subclases entre productos. Segundo, segmentación para asignar clientes a grupos con intereses similares, basado en patrones de compra.

Comienzan con una taxonomía de productos que está disponible en la base de datos. La taxonomía divide los productos en clases, y en subclases. Usan directamente clases y subclases como rasgos de productos y perfiles personales.

Los productos a ser recomendados pueden ser determinados por cálculos de medidas de distancia entre vectores que representan preferencias personales y vectores que representan productos.

Usan el filtrado contextual como fundamento junto con ideas de filtrado colaborativo.

Los datos de entrada consisten de descripciones de aproximadamente 300,000 productos, pero no todos los productos son elegibles para ser recomendados, se evita recomendar tabaco, productos de salud, y otras clases inapropiadas. Todos los clientes en la base de datos son segmentados basados en vectores de gasto normalizados construidos usando su comportamiento de compras previas. Cada cliente es asignado a un solo segmento.

Esta lista específica del cluster pasa por un “motor de coincidencias”, el cual ordena esta lista de productos de acuerdo al atractivo esperado según el objetivo del cliente.

La lista de recomendaciones personalizadas, comprende de 10 a 20 productos con las calificaciones más altas, y regresa al dispositivo durante la siguiente conexión con el servidor. La lista de recomendación contendrá productos previamente no comprados por el cliente.

Los productos se dividen en 99 clases, cada clase tiene alrededor de 100 subclases, generando un total de 2,302 subclases.

El sistema de recomendaciones busca determinar productos los cuales son las mejores coincidencias para este perfil de gasto del cliente. Por lo cual se construye un perfil o vector para cada cliente que representa los intereses del cliente a través del rango de atributos.

Se usa el nivel subclase de la taxonomía de productos para construir este espacio de atributos. Este define un espacio con dimensión de 2,302 en el cual la coincidencia de productos para los clientes.

Se convierte el gasto absoluto para cada subclase a gasto fraccional con el fin de normalizar el gasto total del cliente en el periodo.

Las calificaciones de los gastos normalizados fraccionales de los intereses, normalizadas como entradas iguales para 1 implicación en nivel promedio de interés en una subclase relativa para el resto de los clientes.

Desarrollaron una heurística donde cada tipo de relación entre productos de acuerdo a su taxonomía.

El paso final es calificar cada producto candidato para un cliente específico y selecciona la mejor coincidencia. La calificación podría reflejar un grado de similitudes entre el vector de clientes y el vector de productos.

Limitan el número de recomendaciones de cada cliente a un producto por subclase, y 2 productos por clase.

Usan el Intelligent Miner con el algoritmo A priori para extraer asociaciones entre las 99 clases y también entre 2,302 productos a nivel subclase. Los datos de entrada comprenden 8 semanas de transacciones para 8,000 clientes.

Sólo se calcularon asociaciones simples contenidas en artículos simples.

El algoritmo de recomendaciones combina aspectos de filtrado colaborativo y de contexto para calificar nuevos productos de un cliente base en su comportamiento de compra.

(Baragoian C. et al, 2001) se basaron en el trabajo de Lawrence modificando la fórmula para las entradas en el vector de productos, agregando un valor para indicar suprimir una recomendación de un producto que el cliente ya compró

(Schafer et. al, 2002) proponen con perfiles personalizados para alojar información sobre películas, combinado con personalización de recomendaciones, le da pesos a cada recomendación individual. Indica que los sistemas de meta-recomendaciones ayudan a tener un control personalizado sobre los datos recomendados en diferentes fuentes de información, apoyando al usuario en la decisión de la importancia de cada parámetro. Considera una interfaz para hacer consultas dinámicas.

(Velooso, A. et al., 2002) estudian el algoritmo y los datos de las reglas de asociación por el desempeño de un sistema de evaluación experimental de algoritmos diferentes en diferentes bases de datos. Se observan que cada algoritmo tiene diferentes ventajas y desventajas dependiendo de las características de los datos. El algoritmo tiene un mejor desempeño que previos.

(Ding Q. et al, 2002) utilizan un enfoque diferente al convencional, aplica las reglas de asociación a imágenes de sensores remotos (Remote Sensed Imagery RSI), extrayendo patrones de interés y reglas compuestas de imágenes y datos referentes a la tierra como el clima. Esto puede ser importante en la precisión de la agricultura, la planeación de comunidades, descubrimiento de recursos y otras áreas. Sin embargo, el tamaño de las imágenes es muy grande para el proceso de minado, por lo cual se propone derivar reglas de asociación usando el árbol de Peano (Peano Count Tree, P-Tree), esta estructura permite representar imágenes comprimidas.

(Changchien S.W & Lu Z.C., 2001) recomiendan productos en línea. Amazon.com (Linden G. et al, 2003) emplea filtrado colaborativo.

Por otra parte (Paolo Massa, 2003) pone interés en la confianza de la información para mejorar la precisión y aceptación de los usuarios, basándose en filtrado colaborativo, trata el problema del inicio frío que es cuando un usuario nuevo inicia y no se tiene información sobre él.

(Liu B., 2003) modificó un sistema de clasificación para estimar una calificación que indica la probabilidad de que el dato pertenezca a una clase objetivo. Propone una técnica efectiva y eficiente para calcular probabilidades de clase estimadas usando reglas de asociación.

(Shen Y-D. et al, 2003) estudian los ambientes de comercio electrónico en los cuales el conocimiento de preferencias de clientes puede cambiar rápidamente. Se presenta una propuesta para hacer recomendaciones en línea basadas que incluyen datos pasados y recientes, recabados sólo minutos antes.

(Agrawal et. al 2003) explotaron aspectos grupos para identificar la relación entre miembros comunes, aunque no se aplica directamente a recomendaciones, parece una aplicación lógica para estudios futuros.

(Etzioni et. al, 2003) diseñaron un sistema para minimizar precios de compra para boletos de avión, combina reglas de análisis de series de tiempo, Q-learning, y el algoritmo Ripper, creando un algoritmo de minería multiestrategia, observando tendencias en precios y sugiriendo tiempos de compra.

(Liu D. & Shih Y., 2004) ponen énfasis en el valor del tiempo de vida de los clientes, que generalmente se evalúa por frecuencia y montos, desarrollan una metodología que combina toma de decisiones y minería de datos, utilizando segmentación de clientes de acuerdo a la lealtad o el valor de vida, y posteriormente se utilizaron reglas de asociación para proveer recomendaciones a cada grupo específico. Resultando que la metodología es más efectiva para clientes leales. Posteriormente (Liu D. & Shih Y., 2005) y basados en su trabajo anterior proponen dos métodos híbridos que exploran el método con base en pesos y filtrado colaborativo.

(Cokrowijoyo H. et al., 2004) estudian las reglas de asociación de minería en un DW para enfocarse en la medición de datos sumados. Proponen 2 algoritmos proporcionando los datos iniciales para minar con reglas de asociación en un DW concentrándose en la medición de datos agregados.

(Jacobo, J, et al., 2004) presentan una propuesta para completar el diseño conceptual de DW junto con las reglas de asociación de minería de datos, alojando para implementar las reglas de asociación definidas en la fase de modelado conceptual. Implementan las reglas de asociación especificadas en un servidor DW de administración comercial.

(Ding, Q, et al., 2004) hacen una extracción de patrones de interés y reglas de datos compuestos de imágenes y datos referentes a la tierra, que puede ser de importancia precisión de la agricultura, la planeación comunitaria, descubrimiento de recursos y otras áreas.

Proponen derivar reglas de asociación en datos RSI usando la estructura de árbol Peano Count Tree (P-tree). Resultados experimentales mostraron que el algoritmo propuesto es mejor para datos espaciales RSI.

(Demiriz, 2004) propone el sistema e-VZpro, el cual se basa en 2 fases, en la primera se determinan las reglas de asociación, y en la segunda se usa un algoritmo para calificar las recomendaciones que el cliente hace en línea.

(Daly O., 2004) se enfoca a las reglas de excepción, que son aquellas de interés bajo con confianza alta. Se analiza la interconexión entre las reglas de asociación de excepción y las negativas, y se propone una medida para excepciones para evaluar reglas de excepción candidatas.

(Schafer J. B. et al, 2005) mencionan que en un mundo lleno de opciones un sistema de recomendaciones puede ayudar a evaluar y encontrar productos de interés. Propone un clasificador usando información de productos y cliente como entrada, y la salida es cuanto

recomendar un producto a un cliente. Propone su implementación usando estrategias con reglas de inducción, redes neuronales y bayesianas a través de un árbol de decisiones con información del usuario. El modelo es simple, rápido y tan preciso como el método CF, y podría usarse cuando las preferencias cambian lentamente. Indican que combinar clasificadores con CF agrega valor a las recomendaciones.

Por otra parte (Duen-Ren Liu & Ya-Yueh Shih, 2005) usan la demanda de los clientes derivada de la frecuencia de compra de tipos de productos, exploran dos métodos híbridos que combinan CF y la demanda de clientes para mejorar la calidad de las recomendaciones, utilizan productos re-calificados, y sus experimentos indican que la calidad de la recomendación es muy buena.

Regresando a (Schafer, J. B. 2005), dice que la mayoría de los sistemas de recomendaciones se basa en 3 tecnologías: minería de datos, filtrado y recuperación de información (IF), y filtrado colaborativo. Los sistemas que usan minería de datos son comunes en comercio electrónico, los sistemas de filtro y recuperación se usan para satisfacciones efímeras de bases de datos relativamente estáticas. Sistemas híbridos como las herramientas de búsqueda de Google (www.google.com) combina resultados de ambas búsquedas de contenido y recomendaciones colaborativas.

Propone un sistema meta-recomendador de películas basado en sistemas híbridos, el cual permite un control personalizado en la generación de una lista con datos de diferentes fuentes y técnicas. MetaLens usa tecnologías de IF y de CF, para generar calificaciones de muchos sitios de películas. El sistema provee lo mejor de los dos mundos, aloja recomendaciones basadas en conocimiento persistente del usuario, y permite el control de como combinar datos de recomendaciones. Aunque los resultados están dominados por el dominio de películas.

(Zhan, J. et al., 2005) se enfocan al compartimiento de datos entre dos partes involucradas, estudian como compartir datos privados o confidenciales cuando cada parte tiene un conjunto de datos privados, para crear reglas colaborativas de asociación sin desempacar los datos privados de cada parte. Definen un protocolo usando técnicas de encriptación para intercambiar los datos, las partes son tratadas simétricamente: todos participan en la encriptación y en el empaquetamiento computacional.

(Ogasawara M. et al, 2005) usan reglas de asociación para analizar la relación entre 6 estilos de vida (sobrepeso, bebiendo, fumando, alimentos, ejercicio físico y tiempo de sueño), 5 familias de historias médicas (hipertensión, diabetes, enfermedades cardiovasculares, enfermedades cerebrovasculares, y enfermedades de vida), y 6 anormalidades médicas (alta presión arterial, hipercolesterolemia, hipertriglicéridos, alto nivel de azúcar en la sangre, hiperuricemia, y disfunción de vida) usando exámenes médicos de 7 años, obtenidos de 5350 mujeres empleadas en un grupo de edad de 40-49 años. Encontraron que el número de combinaciones derivadas de minería de datos (reglas de asociación) fue mayor que el derivado de métodos convencionales (análisis de regresión logística), también hallaron que las reglas de asociación son mejores para esclarecer combinaciones efectivas de factores de riesgo en términos de estilo de vida de enfermedades.

(Sriphaew K. & Theeramunkong T. 2005) proponen un método para explotar ocurrencias unigramas y bigramas entre documentos, para extraer un conjunto de documentos con tópicos similares, usando reglas de asociación. Hacen tres propuestas: retroalimentación de información (Information retrieval IR), categorización del texto (text categorization TC) y minería de texto (text mining TM). El método aplica el algoritmo del FP_tree y proponen un matriz de referencias para evaluar las reglas de asociación descubiertas.

(Zhan J. et al. 2005) estudian como compartir datos privados o confidenciales sin descomponer sus partes privadas. Definen protocolos usando técnicas de encriptación para intercambio de datos, donde no es necesario enviar todos sus datos, sólo las partes confiables.

Ponen énfasis en la colaboración heterogénea donde cada parte tiene diferentes conjuntos de atributos.

(Berzal F. et al. 2005) Utilizan modelo de clasificación con reglas de asociación con apoyándose en la confianza y el soporte, usando una representación lógica de implicación del tipo \Rightarrow , donde la intersección de ambos conjuntos es vacía. Construyendo métricas diferentes, las cuales son criterios que pueden ser usados para guiar el proceso de descubrimiento de conocimiento de acuerdo a metas particulares y necesidades de un problema dado.

CONCLUSIONES

Una tienda departamental tiene un reto importante debido a la competencia extrema que existe en cuanto a precios y productos, así como la publicidad para el cliente. Que como ya se ha visto, cada cliente tiene diferentes hábitos de consumo y por tanto la manera de comunicarnos con ellos y lo que se le ofrece no puede ser la misma para todos.

Por otra parte la Minería de Datos constituye la herramienta base para soportar las relaciones entre productos, y de cuyos resultados se obtendrán mejores recomendaciones, que conjuntamente con los hábitos del cliente pueden generar conocimiento sobre los gustos y preferencias del cliente.

También existe la comunicación con el cliente, a través de diversos medios con diferentes efectividades, la forma de comunicación así como la frecuencia, para que al final, lo que se le dice al cliente resulte interesante, de otra manera resulta ineficiente o hasta contraproducente. Por esto mismo lo más importante es encontrar los mejores vínculos entre productos, para lo cual la Minería de Datos resulta importantísima como elemento de análisis.

El ofrecer productos de interés para el cliente tiene el fin de incrementar sus compras y por consecuencia incrementar la fidelidad del mismo, que es lo que todo negocio busca.

Este trabajo tiene como objetivo generar recomendaciones personalizadas para cada cliente, apoyándose principalmente en la Minería de Datos para hallar las relaciones entre productos, y de acuerdo con los hábitos únicos de compra. Para lo cual se usa una metodología que se desarrolla

posteriormente, la cual fue tomada del trabajo de “Lawrence R.D. et al” desarrollado en el 2001, y al cual se le hicieron algunas mejoras para el caso que se desarrolla en este trabajo.

En el siguiente capítulo se habla de la información, tanto de la parte tecnológica como la utilidad y explotación de la misma.

Capítulo II

Base de Datos

INTRODUCCIÓN

En el capítulo anterior se habló sobre lo importante y necesario que es el manejo de información, ya que esto ayuda a entender los hábitos de compra de cada cliente. Implica la manipulación de grandes volúmenes de información donde el manejo manual no es posible, por ejemplo en hojas de cálculo o archivos de texto por la gran cantidad de datos, esto hace necesario contar con una base de datos, que sea la fuente de información para conocer los hábitos de compra y para la generación de las recomendaciones personalizadas.

En este capítulo se revisan conceptos referentes a base de datos, desde el diseño, desarrollo e implementación. El capítulo está dividido de la siguiente manera, en la primera parte, se revisan los conceptos básicos de base de datos, se selecciona el tipo de base de datos a usar en el proyecto; en la segunda parte se diseña e implementa la estructura de datos, se determina la plataforma acorde a los recursos que se tienen, el diseño del modelo de datos y las estructuras principales de la base de datos; en la cuarta parte se planean y obtienen los datos requeridos; en la última parte se preparan los datos para llevar a cabo el proceso de la generación de recomendaciones personalizadas, que será fuente de información de la herramienta de Minería de Datos.

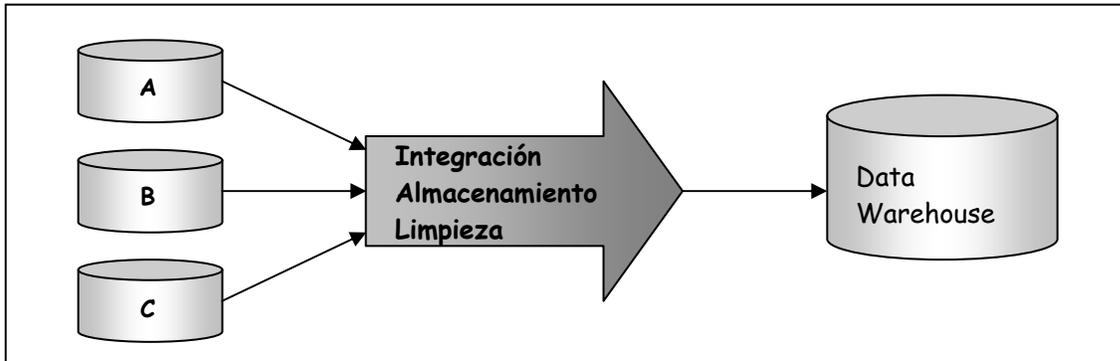
2.1 CONCEPTOS BÁSICOS

Existen bases de datos de diversos tipos que son construidas con diferentes propósitos, desde las que tienen fines operativos hasta las de análisis.

Las bases de datos y aplicaciones basadas en procesamiento tradicional se conoce como procesamiento transaccional en línea (OLTP, On-Line Transaction Processing), que son suficientes para cubrir las necesidades operativas no así para análisis, planeación o toma de decisiones estratégicas. Usualmente los datos de análisis pertenecen a diversas áreas, fuentes externas e incluso que nunca antes hayan sido recopilados. Integrar esta información no es simple y representa un reto enorme en cualquier organización, esta tecnología se llama Data Warehousing (figura 2.1).

El diseño de una base de datos, Data Warehouse (DW), difiere substancialmente del diseño de una base de datos para un sistema transaccional, OLTP (Tabla 2.1).

Una base de datos relacional se diseña con un propósito específico, ya que el propósito de un Data Warehouse difiere de uno de un OLTP, el diseño de sus características difiere enormemente uno del otro.



Tecnología Data Warehousing

Figura 2.1

Base de datos Data Warehouse	Base de datos OLTP
Para análisis de métricas del negocio usando categorías y atributos.	Diseñada para operaciones del negocio en tiempo real.
Optimizada para volúmenes altos, complejos, consultas no predecibles que acceden a muchos renglones de una tabla.	Optimizada para conjuntos de transacciones comunes.
Cargada con consistencia, validación de datos, consultas que no requieren validación en tiempo real.	Optimización de validaciones de entrada de datos durante las transacciones, usa tablas de validación.
Soporta pocos usuarios concurrentes, en relación con OLTP.	Soporta miles de usuarios concurrentes.

Propiedades de las bases de datos DW y OLTP

Tabla 2.1

Un Data Warehouse soporta un sistema OLTP, acumula información del OLTP, y provee servicios que podrían complicar las operaciones del OLTP si fueran hechas en esa base de datos. El DW guarda información histórica y su propósito es el manejo de grandes volúmenes de información a través de consultas.

Si los datos son guardados sólo para preservarlos, no los hace viables para ser usados por analistas o tomadores de decisiones. Si los datos son acumulados y organizados entonces pueden ser usados y explotados con fines de análisis. Las consultas en los OLTP pueden llegar a ser muy

complejas computacionalmente hablando, mientras que un DW puede ser muy eficiente, sin pasar por el OLTP.

Un DW es un repositorio donde se guardan datos de diversas fuentes de información, que agrega e integra información, los datos se modelan con una estructura de datos multidimensional, en donde cada dimensión corresponde a un atributo o a un conjunto de atributos de los hechos, esto hace que un DW sea adecuado para el procesamiento analítico en línea (On-line Analytical Processing, OLAP), permitiendo un análisis multidimensional para la presentación en diferentes niveles de abstracción.

La Minería de Datos difiere de la tecnología OLAP, en OLAP se usan la herramienta para obtener información agregada a partir de información detallada, puede usarse para comprobar patrones e hipótesis, es un proceso deductivo, mientras que la minería de datos usa los datos para encontrar patrones, es un proceso inductivo.

La Minería de Datos puede utilizar como fuente de información un DW pero existen otras posibilidades que pueden resultar mejores para la explotación de la información.

Las aplicaciones de base de datos se pueden categorizar según, Scalzo 2003, como se muestra a continuación (Tabla 2.2):

	OLTP	ODS	OLAP	DM/DW
Enfoque del sistema	Operacional	Operacional/Táctico	Táctico	Táctico/Estratégico
Herramientas de usuarios finales	Cliente/Servidor o Web	Cliente/Servidor o Web	Cliente/Servidor	Cliente/Servidor o Web
Tecnología DB	Relacional	Relacional	Cúbica	Relacional
Cálculo de transacciones	Grande	Mediano	Pequeño	Pequeño
Tamaño de las transacciones	Pequeño	Mediano	Mediano	Grande
Tiempo de las transacciones	Pequeño	Mediano	Mediano	Grande
Modelación de datos	Tradicional ERD	Tradicional ERD	N/A	Dimensional
Normalización	3-5 NF1	3 NF	N/A	0 NF

Categorización de Base de datos

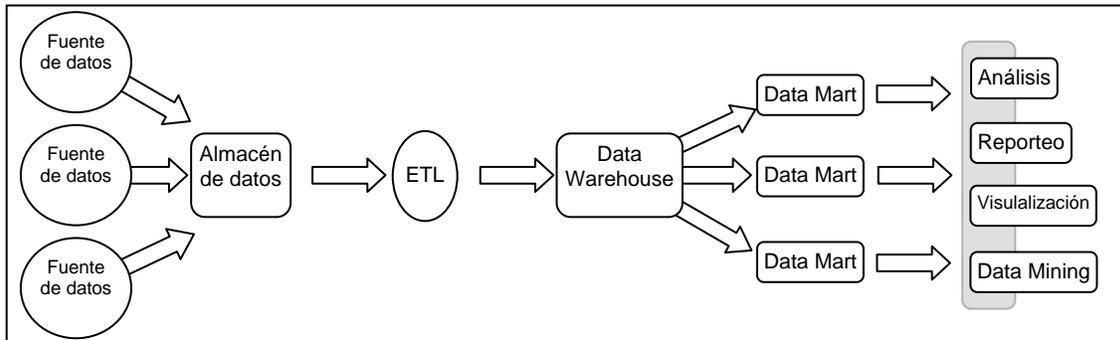
Tabla 2.2

- Procesos transaccionales en línea (On-line Transaction Processing, OLTP)
- Almacenes de datos operacionales (Operational Data Store, ODS)
- Procesos analíticos en línea (On-line Analytical Processing, OLAP)

- Data Mart/Data Warehouse (DM/DW)

Según McCullough 2003, el flujo de información en las bases de datos se puede considerar como una “Cadena de Suministro de Información”.

Un proceso importante durante la construcción y mantenimiento de bases de datos es el ETL (Extraction, Transformation, Load), que construye con características específicas en cada caso, la carga y mantenimiento de los datos (Figura 2.2).



Carga y mantenimiento de datos

Figura 2.2

Por otra parte, Rizzi 2001, habla sobre el valor de la información de acuerdo al costo de la misma (Figura 2.3).

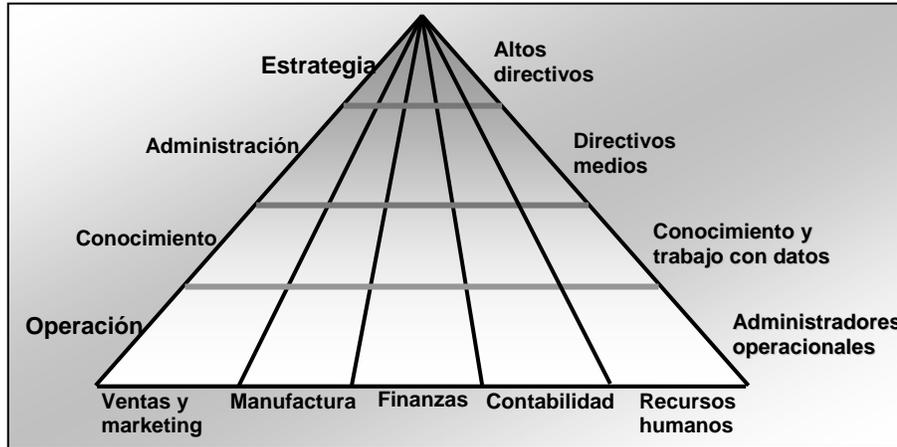
- La información es un recurso de la empresa como capital, primera materia, plantas y personas; por esta razón es que tiene un costo.
- Por este costo resulta importante entender el valor de la información.



Valor de la información por costo

Figura 2.3

También habla sobre las diferentes clases de sistemas de información y su valor de acuerdo al nivel de utilización y al nivel de utilización ende la jerarquía de cargos (Figura 2.4).



Valor de la información por utilización

Figura 2.4

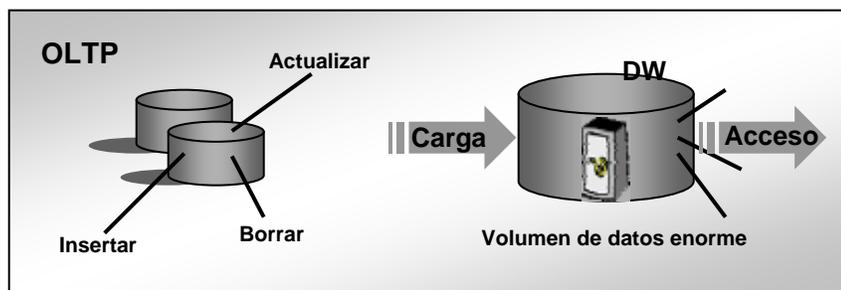
La información también tiene un valor por su evolución temporal, es decir por su desarrollo a través del tiempo o historia, lo cual se refleja considerablemente con las diferencias entre un DW y un OLTP (Tabla 2.3).

OLTP	DW
Valores actuales	Instantáneo
Contenido histórico restringido	Contenido histórico rico
Frecuentemente no se incluye tiempo en la llave	El tiempo se incluye en la llave
Los datos son actualizados	Los instantes no pueden ser actualizados

Valor de la información en el tiempo

Tabla 2.3

Que también hacen una diferencia el volumen de información acumulado en el tiempo, y la cantidad de recursos utilizados para su almacenamiento (Figura 2.5).



Valor de la información en el tiempo por volumen

Figura 2.5

Las aplicaciones de BI (Business Inteligent), toma de decisiones y CRM (Customer Relationship Managment) utilizan información registrada en las bases de datos, pero éstas pueden tener formatos diferentes y pueden ser no fácilmente explotables, pero constituyen almacenes de datos donde las aplicaciones podrían extraer libremente información.

A continuación se explica detalladamente las características de un Data Warehouse y un Data Mart.

Data Warehouse (DW).

Un DW es un almacén inmediato de datos extraídos de aplicaciones en producción y que agrupan una copia de los datos vitales de la empresa para periodos largos de tiempo. Puede contener pedidos, facturación, producción, prospectos, ventas, datos crediticios, etc. Se utiliza con fines de reintegración, análisis, conocer la actividad o determinar tendencias para la toma de decisiones, con las características siguientes:

- Integración, agregar aquellos datos completos y directamente explotables.
- No volatilidad, es decir que los datos no pueden ser suprimidos después de su explotación.
- Histórico. Los datos se sitúan en el tiempo, con el fin de estudiar su evolución.

El DW corresponde a un proyecto donde se manejen grandes volúmenes de información.

Data Mart (DM).

Un DM está reservado para volúmenes de datos pequeños, es dirigido a una actividad estratégica o un departamento específico, y la información tiene el fin de ser explotada directamente. Un DM es una aglomeración de un DW más enfocada a un objetivo o tema.

De acuerdo a Gartner 2001, las diferencias entre un DW y un DM son las siguientes (Tabla 2.4).

Para una organización es difícil integrar toda la información en el formato estrella, este formato es una forma de representación y organización de la información que posibilita determinar cual es el objeto central de análisis y cual es la información adyacente o variables de análisis, propio de un DW, por lo cual resulta más eficiente en algunos casos construir Data Marts para cada área temática de la organización, donde cada uno tiene medidas y dimensiones propias.

Un Data Mart contiene datos de DW integrados para soportar los requerimientos específicos del negocio, en suma los principales propósitos de un Data Mart son los dos siguientes:

- Almacenar información pre-agregada.
- Proveer acceso rápido para necesidades específicas de análisis.

Un Data Mart se implementa como una extensión de un DW, no es un sustituto, y es construido para un propósito específico, esto hace que sea una fuente de datos idónea para la Minería de Datos, que analiza los datos relevantes específicos y extrae patrones ocultos o evidentes, de esto se habla a fondo en el capítulo 3.

Aspectos	Data Warehouse	Data Mart
Alcance	Aplicación neutral Centralizada, compartido Cross-LOB/empresa Con arquitectura	Requerimientos específicos de la aplicación. LOB, departamento o área usuaria Orientación procesos de negocio Múltiples bases de datos con datos redundantes.
Perspectiva de datos	Detalle histórico de datos Algunos agregados Ligeramente desnormalizada	Detallada (alguna historia) Agregada Altamente desnormalizada
Tema	Múltiples áreas temáticas	Una sólo área temática parcial. Múltiples áreas temáticas parciales. Fuentes operacionales instantáneas
Fuentes de datos	Muchos Operacional, datos externos	Pocos Operacionales, datos externos Base de datos OLTP instantáneas. Extractos de datos "piratas"
Tiempo implementación	9 a 18 meses para la primera etapa (2 o 3 áreas temáticas) Múltiple implementación de etapas.	4 a 12 meses.
Características	Flexible, extensible Durable/estratégica Orientación de datos	Restringida, no extensible Vida corta/táctica Orientación de proyecto

Diferencias entre DW y DM

Tabla 2.4

2.2 ARQUITECTURA DE DATOS

La información surge en las terminales punto de venta, que es en donde se realizan las transacciones, posteriormente estos datos se procesan y son almacenados en bases de datos, las cuales tienen como propósito administrar las operaciones de compra que realizan los clientes. En la tienda departamental de la que se está hablando, existe una infraestructura de BI (Business Intelligence) para almacenamiento de datos, lo que significa que se tiene la ventaja de que la adquisición de datos no parte desde cero.

Se cuenta con un ambiente de DW en producción, con lo cual, la extracción de datos se dirige totalmente hacia éste. Dicho DW es corporativo y cuenta con dos grandes módulos de interés, uno destinado para el área temática de mercaderías y otro para el área temática de crédito. Cada uno tiene propósitos diferentes, mientras uno de ellos está dirigido al análisis de productos, el otro está dirigido a la administración del crédito de los clientes. Pero en ninguno de los dos casos se tiene como objetivo el análisis de las compras del cliente, integrando tanto la parte de productos como la

información crediticia y los datos demográficos, que son analizados en el área de CRM (Customer Relationship Management).

De aquí surge la necesidad de construir un modelo de datos lógico orientado a satisfacer las necesidades generales del área temática de CRM. Este modelo representa los elementos de datos y sus relaciones combinando información de productos e información crediticia, así como otros elementos propios de CRM.

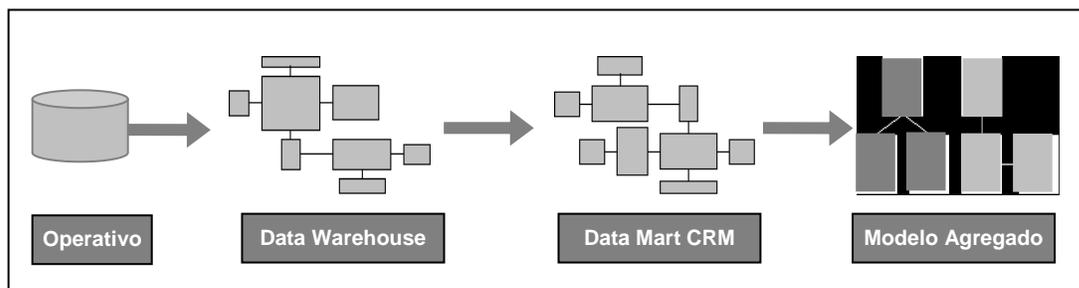
Debido a que el modelo de datos está orientado a una entidad organizacional específica con un propósito concreto, es considerado como un Data Mart, al cual se hará referencia de ahora en adelante como el "Data Mart CRM". La implementación es de suma importancia, ya que éste es la fuente de información para la herramienta de Minería de Datos y por consiguiente para la generación de recomendaciones; los beneficios adicionales generados con su implementación son la independencia del resto de los usuarios (en términos de concurrencia y consumo de recursos), la autonomía que proporciona al área de CRM (en términos de autoservicio de información), la descarga de trabajo al área de sistemas (en términos de reproceso continuo de extracciones equivalentes) y la facilidad para combinar información inter-departamental bien consolidada (en términos de simplicidad y precisión), pero principalmente es la integración de la información del cliente y su historial de compra a nivel producto.

Para satisfacer las necesidades de información de este proyecto es necesario construir un modelo agregado que cumpla con los requerimientos de extracción de información para la generación de recomendaciones (figura 2.6).

La información agregada hace que la explotación de la misma sea simple y que se pueda tener riqueza de contenido, por lo cual se decide integrar el modelo agregado, que es fuente de información para el algoritmo de Minería de Datos, al Data Mart CRM como un área temática adicional. De esta forma puede ser además explotado con las herramientas de consulta con las que ya se cuenta, para analizar y buscar la información que se encuentra en el DW Corporativo.

Son 4 los niveles de información involucrados (Figura 2.6):

- 1) El operativo; 2) El Data Warehouse Corporativo (DWC); 3) El Data Mart CRM (DM-CRM) y 4) El Modelo Agregado de GRP (MA-GRP).



Generación del Modelo Agregado

Figura 2.6

2.2.1 Plataforma

La plataforma en la que reside el DW es Redbrick, también para el Data Mart de CRM, esto debido a la compatibilidad de plataformas y a costos, no es necesario adquirir una nueva base de datos y servidor, el costo se reduce a la utilización de disco duro del servidor actual y un mantenimiento conjunto.

La parte más relevante del proceso de generación de recomendaciones, es la utilización de un algoritmo de Minería de Datos, que será explicado a detalle en el siguiente capítulo, la empresa tiene una software de Minería de Datos llamado Intelligent Miner de IBM con una base de datos DB2, que serán las herramientas usadas en el proceso de generación de asociaciones entre productos.

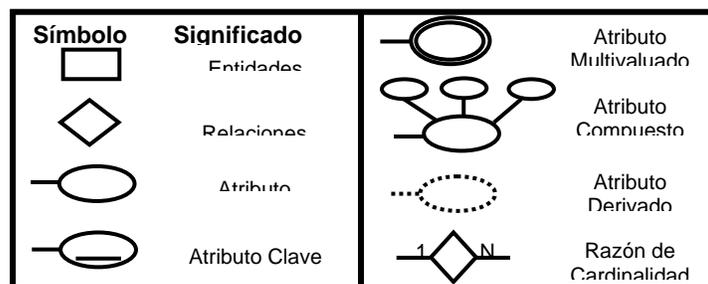
Ya existen procesos de extracción, transformación y carga de información para poblar el DW corporativo, después de construir el Data Mart de CRM, deben construirse procesos de extracción, transformación y carga de información, así también debe hacerse lo mismo para poblar las tablas del Minero de Datos necesarias y definidas por el algoritmo elegido, que se explica en el capítulo siguiente.

2.2.2 Modelo de datos de negocio – ER

El modelo Entidad Relación ER se usa para construir bases de datos relacionales simples y que optimizan las consultas, minimizando el número de tablas y sus relaciones. Es un proceso para diseñar y crear los elementos que componen un sistema de base de datos. ER se basa en la percepción de la realidad, describe los datos como entidades, relaciones y atributos.

Existen diferentes modelos de organización de bases de datos, en este trabajo se trata el modelo ER, debido a que se utiliza para la implementación y no se profundiza en el proceso de almacenamiento físico de los datos.

Con el objetivo de facilitar el proceso de diseño de la base de datos, el modelo ER, utiliza la simbología que se muestra en la figura 2.7.



Notación del modelo ER
Figura 2.7

El modelo de datos es una representación abstracta de los datos, que define la manera en que son organizados y relacionados, por ejemplo, gráficas, fórmulas matemáticas, tablas, etc.

En el modelo relacional todos los datos se representan en tablas, las tablas están compuestas por filas y columnas; las filas sólo se ordenan si se le indica a la base de datos que lo haga, de no ser así, el orden será arbitrario y puede cambiar en caso de tratarse de una base de datos dinámica; cada tabla tiene una clave primaria, un identificador único, compuesto por una o más columnas; para establecer una relación entre dos tablas es necesario incluir una o más columnas que representen a la llave primaria, en la otra(s) tabla(s), a esta columna se le llama clave secundaria.

Un buen diseño de base de datos debe reflejar la estructura del problema en el mundo real, ser capaz de representar todos los datos esperados, incluso con el paso del tiempo, evitar el almacenamiento de información redundante, proporcionar un acceso eficaz a los datos, mantener la integridad de los datos a lo largo del tiempo y ser claro, coherente y de fácil comprensión.

El proceso de diseño en el modelo ER debe identificar las entidades que debe presentar la base de datos, determinar las cardinalidades de las relaciones entre las distintas entidades y clasificar estas relaciones en una de los siguientes tipos:

1. **Uno a uno**, por ejemplo, un cliente sólo tiene un género.
2. **Uno a muchos**, por ejemplo, una línea de crédito puede tener un cliente titular y uno o más adicionales.
3. **Muchos a muchos**, por ejemplo, un cliente puede comprar un varios departamentos y en cada departamento compran varios clientes.

Es necesario dibujar el diagrama Entidad Relación, así como determinar los atributos de cada entidad y definir la llave primaria (única) de cada entidad, sin olvidar que las entidades entre las que hay una interrelación uno a uno se deben fusionar en una sola entidad. Posteriormente, cada una de las entidades restante se convierte en una tabla con una llave primaria y una serie de atributos, de los cuales algunos pueden ser claves secundarias. Las relaciones uno a muchos se transforman en atributo y clave secundaria de la tabla que representa a la entidad situada del lado de la relación correspondiente a muchos. Las relaciones muchos a muchos entre dos entidades pasan a ser una tercera tabla con llaves secundarias procedentes de ambas entidades. Estas claves secundarias deberán formar parte de la clave primaria de la tabla en la que se convierte la relación, cuando corresponda. Hay una serie de herramientas disponibles que pueden automatizar el proceso de conversión de un modelo ER en un esquema de base de datos, es decir pasar del diseño a la creación de la base de datos.

El modelo de datos desarrollado en este trabajo puede consultarse en el Anexo 2, con las estructuras y sus relaciones.

2.2.3 Estructuras principales

El Data Mart de CRM fusiona información de mercadería y crédito, dentro del DM también se creó un área temática para el apoyo en la generación de recomendaciones personalizadas con información nueva, derivada o externa incorporada al modelo de datos. A continuación se describen las principales estructuras de información creadas (que pueden ser consultadas en el Anexo 2).

- **Rango Moda**

Se generó una matriz de marcas, esta matriz clasifica las marcas por departamento en un rango de moda de 5 niveles que se mueve de lo muy clásico a lo muy contemporáneo, de esta manera cada producto se clasifica automáticamente de acuerdo a su departamento y a su marca (ya posicionado en la matriz de marcas). Así, se tiene una clasificación detallada de cada producto, por ejemplo ropa clásica, zapatos clásicos, accesorios clásicos, muebles clásicos. De tal manera que un cliente cuya marca favorita es una clásica, sólo tenga recomendaciones de marcas muy clásicas o clásicas.

- **Rango Precio**

La matriz de marcas también clasifica por departamento y marca dentro de un rango de precios de 6 niveles, que se mueve de Bajo hasta Bridge Alto, de esta manera los productos se clasifican automáticamente por su departamento y marca (previamente ubicados en la matriz de marcas por rango de precios). De esta manera, se tiene automáticamente una clasificación detallada de cada producto, por ejemplo un vestido puede ir de valor bajo a bridge alto. De tal manera que un cliente cuya marca favorita es una de valor alto, sólo tenga recomendaciones de productos de precios altos.

- **Matriz de Marcas**

Es una entidad atributiva en la que se identifica la ubicación en precio y moda de las diferentes marcas clasificadas por departamento. Es decir que para cada marca por departamento se tiene identificada dentro de un rango de moda y precio. Esta entidad es clave para el refinamiento de las listas de recomendaciones. Clasifica a las marcas dentro de un rango de moda y precio, que son únicos dentro de cada departamento, y que puede variar en departamentos diferentes una misma marca.

- **Marca Afín**

Es una entidad en la que de manera explícita se presentan todas las marcas afines que una cierta marca puede tener. Se tiene un conjunto de marcas que son similares en cuanto a moda y precio, y que pueden ser opcionales a la marca favorita del cliente, ya que son análogas y muy probablemente del gusto y capacidad de compra del cliente. Este dato es clave para el mejoramiento de las listas de recomendaciones, ya que además de recomendar un producto se agrega la marca favorita del cliente o como segunda opción una marca afín a los gustos y capacidades de compra del cliente.

- **Tarjeta Alta Diversidad**

Es una entidad de tarjeta en la que se registran únicamente aquellas tarjetas que cumplen con los requisitos de alta diversidad, es decir, el número de tipos de productos comprados y el porcentaje de gasto relativo máximo. Esta entidad es clave para la determinación de las reglas de asociación con el algoritmo de Minería de Datos, ya que se filtran sólo aquellas tarjetas que contengan suficiente información, como para encontrar relaciones importantes entre clases de productos. El propósito es eliminar aquellos clientes cuyas compras han sido de muy pocos productos y que por lo tanto no aportan mucha información al algoritmo de asociaciones.

- **Tarjeta Marca Preferida**

Es una entidad de tarjeta en la que se registra únicamente la marca de mayor consumo observada para cada tarjeta que participa en el análisis. Es decir, aquellas marcas que son compradas más frecuentemente por cada cliente. Es clave para el perfeccionamiento de las listas de recomendaciones, ya que la recomendación de un producto va acompañada con la marca favorita del cliente, en caso de ser posible.

- **Prorrateo Compra**

Es una entidad derivada en la cual se manifiesta el efecto de prorrateo que se produce al combinar las mercancías adquiridas en una transacción, junto con las formas de pago empleadas para pagar. Hay casos en que los clientes utilizan más de una forma de pago en la adquisición de un producto, puede ser que un porcentaje del monto lo paguen en efectivo o con una tarjeta de crédito o con un cupón. Por ejemplo si el producto cuesta \$100 y la mitad se paga en efectivo y la otra con la tarjeta de crédito, en la base de datos fuente se registra que el cliente compró un producto de \$50 con la tarjeta, cuando en realidad \$50 se refiere al pago no al producto. Para no perder de vista estos casos, se genera un proceso donde se asigna que el 50% del producto se pagó con la tarjeta de crédito y no que el producto costó \$50. El procedimiento implica un cruce exhaustivo entre productos y tipos de pago. Esta entidad es la fuente de información de más bajo detalle y a

partir de ella se construyen varias agregaciones de datos que dan origen a varias piezas de información utilizadas para la generación de recomendaciones personalizadas.

- **Vector Tarjeta Marca**

Es una entidad derivada, es decir que no existe en las fuentes de información, si no que se calcula con un proceso, presenta datos de consumo relativos a las preferencias de marca para cada tarjetahabiente, con base en los montos y frecuencias de compra de cada cliente.

- **Vector Tarjeta Clase**

Como ya se ha mencionado cada departamento tiene una clasificación de productos dentro de la jerarquía llamada clase, dicha clase indica de manera general el tipo del producto, pero sin llegar a detalles específicos como el color o la talla. Algunos ejemplos son: para el departamento Damas: vestido, falda, conjunto; para el departamento Caballeros: traje, saco, pantalón.

Es una entidad derivada de tipo asociativa, presenta datos de consumo relativos a las preferencias de clase para cada tarjetahabiente. Es una entidad muy relevante, ya que con ésta se determinan los hábitos de compra y la calificación de propensión que da origen a las listas de recomendaciones, se usa para determinar el tipo de cosas que compra cada cliente a través del tiempo, que posteriormente ayuda a crear una recomendación de acuerdo a cada caso.

- **Vector Clase Clase**

Las recomendaciones se basan en los hábitos de compra generales, es decir en la asociación de productos que se genera a partir de la compra conjunta de todos los clientes, y cuyas relaciones serán encontradas a través de un algoritmo de Minería de Datos. Esta entidad asocia información relativa al consumo y a las afinidades de cada clase hacia otras clases. Es clave para la determinación de la calificación de propensión, cuyo valor será encontrado por el algoritmo de Asociaciones, contiene la relación entre productos a nivel jerárquico de clase. Por ejemplo la clase Vino Tinto está asociada a la clase Queso Fuerte, y la clase Cerveza está asociada a la clase Queso Suave.

- **Departamento Unificado**

Como ya se mencionó, una de las fuentes de información es el DW, cuyo propósito es el manejo de los artículos, por lo cual un departamento que un cliente ve como único (por ejemplo Gourmet), administrativamente está compuesto de diversos departamentos (Gourmet tipo1, Gourmet tipo2 y Gourmet tipo3). Para unirlos y verlos como uno solo se crea una entidad que los clasifica a un conjunto de la misma naturaleza como únicos (los tres tipos en uno llamado Gourmet). Por otra parte se suman las compras de cada cliente en uno sólo, por ejemplo una persona que compró un Vino Tinto en el departamento Gourmet tipo1 y otro Vino Tinto en el departamento Gourmet tipo2,

en la información original el cliente compró en dos departamentos diferentes departamentos, con este dato la información se agrega a un departamento y aparece que el cliente compró dos vinos tintos en el mismo departamento. Funciona como conmutador para la consolidación de variables de consumo, así como para enriquecer los datos fuentes del algoritmo de Minería de Datos.

- **Clase Resuelta**

Dentro de las clasificaciones de productos, cada departamento unificado tiene sus propias clases, por ejemplo para el departamento Vinos y Licores la clase 1 es Vino Tinto, mientras que para el departamento Sonido y TV la clase 1 es Televisión. Para evitar el cálculo de dicha la asociación, se genera previamente con el objetivo de agilizar el cálculo de la recomendación durante el proceso.

Es una entidad en la que de manera explícita se incluyen las descripciones concatenadas del departamento unificado y de la clase, su creación obedece a consideraciones de desempeño de procesos en el servidor.

- **Estilo Alto Valor**

Es una entidad atributiva del estilo en la que se registran únicamente aquellos estilos que cumplen con el requisito de alto valor, es decir, el importe transaccional promedio al ser adquirido. Esta entidad es clave para la determinación de las reglas de asociación de Minería de Datos entre clases, ya que elimina aquellos artículos de bajo valor y que no aportan información útil, como por ejemplo los dulces a granel, galletas o panes que se venden por pieza.

- **Estilo Unificado Resuelto**

Es una entidad en la que de manera explícita se incluyen las descripciones del departamento unificado, la clase, la subclase y el estilo. Su diseño obedece a consideraciones de desempeño.

- **Estilo Comprado**

Es una entidad derivada en la cual se presenta una alternativa del detalle de la transacción consolidada a nivel estilo. Esta agregación es empleada para determinar los estilos de alto valor de manera eficiente, es decir aquellos que pueden enriquecer las asociaciones de productos.

2.3 OBTENCIÓN DE DATOS

El modelo de datos proporciona la estructura necesaria, en relación a las variables, para llevar a cabo la explotación mediante la Minería de Datos, pero aún es necesario desarrollar los procesos y obtener los datos. Su obtención implica las etapas de identificación, colección, filtrado y agregación de acuerdo al formato requerido por el modelo de datos y las técnicas de minería que pueden ser utilizadas.

A continuación se listan las características de la información con la que se desarrolla el proceso.

Fuente de información

Data Warehouse.

Plataforma: Redbrick.

Periodo de análisis

Enero 2005 - Diciembre 2005.

Características de los clientes

Poseen una tarjeta de crédito de la tienda.

Realizaron al menos una compra en el periodo de análisis.

Nivel de detalle

Transacción (Tkt) – Por artículo.

Tarjeta de Crédito.

Nivel de agregación

A nivel Tarjeta de Crédito.

A nivel Clase.

A nivel Trimestre.

Volúmenes de información

La tabla 2.5 muestra el volumen de información que se utiliza para el desarrollo de los procesos.

Pieza de información	Ocurrencias
Tarjetas participantes	332,470
Transacciones	6,951,258
Departamentos	241
Clases	1,815
Subclases	9,091
Estilos	1,177,051
Marcas	7,804

Volumen de información

Tabla 2.5

2.4 PREPARACIÓN DE DATOS

La extracción y preparación de datos es el proceso de coleccionar datos de diversas fuentes para moverlos a una base de datos. Extracción de datos dentro de un medio ambiente DW es un proceso selectivo para importar información relevante de decisión en el DW.

Es más que copiar datos de una base de datos a otra. Dependiendo de la técnica este proceso también se refiere a Pulling (extracción de datos) o Pushing (Propagación de datos).

La transformación de datos usualmente envuelve resolver código con tablas de mapeo, por ejemplo cambiar:

- ✓ Masculino por → 0
- ✓ Femenino por → 1

Esto implica cambiar el valor de reglas de negocio escondidas en datos, las relaciones se ajustan al análisis de dominio y ocurren transformaciones a través del proceso de población de información, usualmente implica el desarrollo de varios procesos.

La limpieza de datos asegura que el DM tendrá información válida, útil y con significado, así como estandariza datos, se tienen dos puntos importantes en la preparación de datos:

Agregación de datos: cambia el nivel de granularidad en la información.

Suma de datos: Agrega valor a ciertos grupos de información.

La preparación de información para Minado de Datos es usualmente una tarea que consume mucho tiempo, frecuentemente el minado por sí mismo requiere menos esfuerzo en el desarrollo de tareas y tiempo. El camino óptimo para el preprocesamiento de datos de minería depende de la tecnología usada, de las habilidades, del volumen de datos a ser procesados y de la frecuencia de actualización.

Una vez que el modelo de datos ha sido poblado es necesario asegurarse de que la información recibida cumple con integridad, exactitud y relevancia, para verificarlo pueden revisarse estos puntos:

- Inspección visual: Exploración de la información para la detección de posibles distribuciones inverosímiles.
- Identificación de inconsistencias: Corrección de errores, manejo de casos extremos y de información faltante.
- Selección final de variables: Eliminación de variables dependientes o altamente correlacionadas.

CONCLUSIONES

En el transcurso del capítulo se revisó la parte de la ingeniería de información, desde crear una base de datos óptima y congruente con el objetivo del proyecto, así como el diseño del modelo de

datos y la construcción del Data Mart funcional y acorde con los procesos para la generación de recomendaciones personalizadas, y por otra parte con la herramienta de Minería de Datos, con los recursos de software y hardware con que se cuenta. También se señalan las estructuras de datos principales, que ayudaran de manera determinante encontrar mejores relaciones durante el minado de datos con la herramienta dedicada a este propósito.

Se señala que la preparación de información es de vital importancia para el desarrollo de la metodología y ayuda a que los procesos y la generación de información sean más simples, así como que los procesos fluyan de mejor manera, poniendo especial énfasis en la información requerida para alimentar el algoritmo de Asociaciones durante el minado de datos, lo cual se expone con profundidad de detalle en el capítulo siguiente.

Capítulo III

Diseño y Desarrollo del Modelo

INTRODUCCIÓN

Una vez que se cuenta con la infraestructura de información, se puede desarrollar los procesos que llevan al resultado deseado, es decir una lista de recomendaciones para cada cliente, el proceso más importante de todos es el relacionado con la Minería de Datos, que se desarrolla a lo largo de este capítulo.

La primera parte explica la Minería de Datos, los algoritmos que existen, las aplicaciones de los mismos y la elección del algoritmo que se aplica, en la segunda parte expone el proceso para la generación de las recomendaciones personalizadas, en la tercera parte se hace un análisis de información sobre los datos usados en el proceso, la cuarta parte trata sobre la aplicación del algoritmo para la obtención de las reglas de asociaciones de productos, así como la distinción entre tipos de vínculos, la quinta parte habla sobre la construcción de perfiles de consumo o hábitos de compra, relacionados tanto con productos como con marcas, en la sección seis se desarrolla y aplica el algoritmo de Minería de Datos, en la séptima parte se sistematiza el proceso, y por último se mide la respuesta del modelo, y se exponen algunos ejemplos de listas de recomendaciones.

3.1 MINERÍA DE DATOS

El análisis tradicional de datos se ha convertido en insuficiente en nuestros días, ahora son indispensables métodos asistidos por computadora, en particular aquellos de Minería de Datos, así como también otras técnicas relacionadas al descubrimiento del conocimiento en bases de datos, con análisis de datos inteligente.

La información histórica es útil para explicar el pasado, entender el presente y predecir el futuro. El conocimiento que puede inferirse de los datos y la utilización del mismo es muy importante en el desempeño de una organización.

Según Witten & Frank (2000) la Minería de Datos es el proceso de extraer conocimiento útil y comprensible, previamente conocido, desde grandes cantidades de datos almacenados en distintos formatos.

Baragoin et. al. (2001), dicen que la Minería de Datos es la exploración y análisis de grandes volúmenes de datos con procedimientos automáticos para la búsqueda de nueva información previamente desconocida, que sea relevante para el negocio.

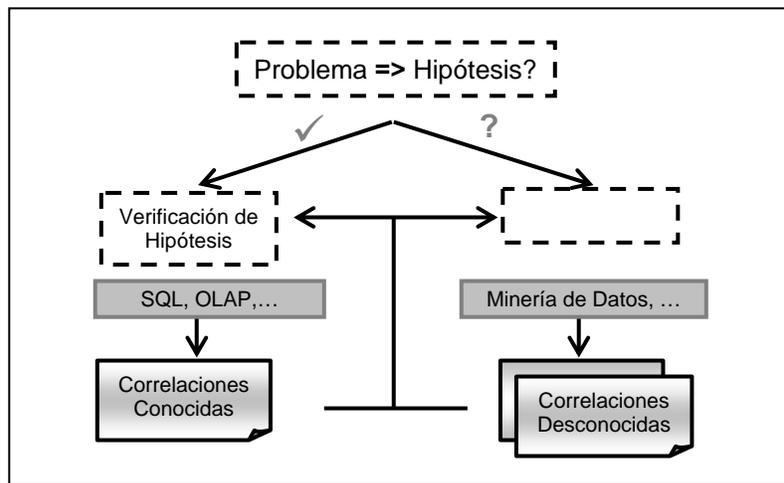
Para Hernández et. al. (2004), la tarea fundamental de la Minería de Datos es encontrar modelos inteligibles a partir de los datos. Dos son los retos, por un lado trabajar con grandes volúmenes de

datos y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. El objetivo de la Minería de Datos es convertir datos en conocimiento.

Ólafsson (2005), dice que la Minería de Datos es un campo interdisciplinario y que encaja forzosamente tanto en la estadística como en el aprendizaje tecnológico.

La Minería de Datos es un área multidisciplinaria, cuyo propósito es generar conocimiento previamente ignorado u oculto, a través de la explotación de grandes volúmenes de información.

La Minería de Datos genera respuestas, sin que previamente se deba hacer una pregunta específica, a diferencia del enfoque SQL (Structured Query Language) u OLAP (On Line Analytical Processing), lo cual se muestra en la figura 3.1.



Enfoque SQL, OLAP y Minería de Datos

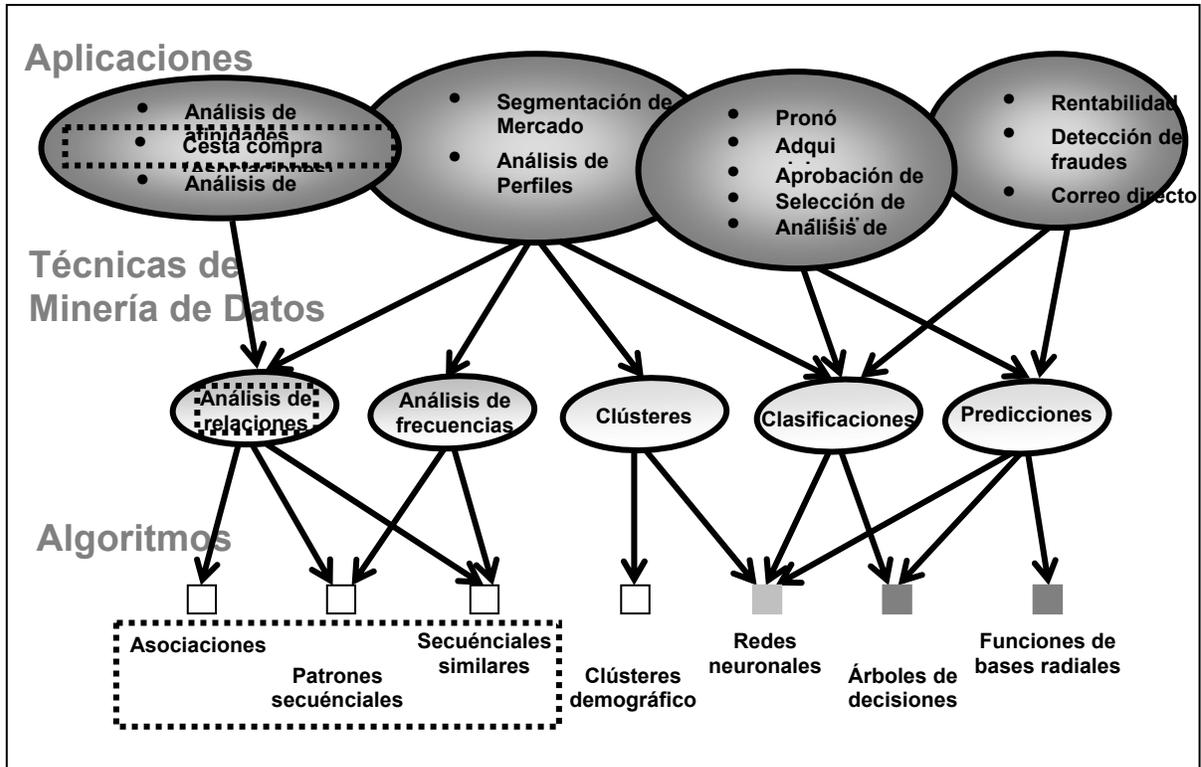
Figura 3.1

La Minería de Datos puede emplearse en diversos tipos de aplicaciones, es usual combinar diferentes técnicas de minado para un propósito particular. La figura 3.2 muestra algunos tipos de aplicaciones, con una serie de situaciones donde la minería se ha usado y una serie de aplicaciones que van desde la segmentación de clientes en tiendas a análisis de riesgos y fraudes en bancos o instituciones financieras.

El problema que aquí se está desarrollando es en una tienda departamental, donde hay clientes a los que se desea hacer recomendaciones personalizadas de productos, que es un caso claro de un análisis de la cesta de compra (Market Basket Analysis), para este tipo de problema se pueden utilizar reglas de asociación, patrones secuenciales o secuencias similares (Figura 3.2), pero el algoritmo a utilizar se define en la sección cuatro.

La metodología utilizada para el desarrollo de este proyecto en Minería de Datos es la metodología genérica, el ciclo de vida de un proyecto de minería consiste de 6 fases: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación, despliegue y generación de hipótesis; estas no que no son rígidas ni secuenciales y usualmente se repiten

varias veces (Figura 3.3). Siempre es necesario regresar y avanzar a través de las diferentes fases. En la figura 3.3 las flechas indican las frecuencias y las dependencias más importantes entre cada una de las seis etapas. El círculo externo representa el ciclo natural de la Minería de Datos.



Aplicaciones, Técnicas y Algoritmos de Minería de Datos

Figura 3.2

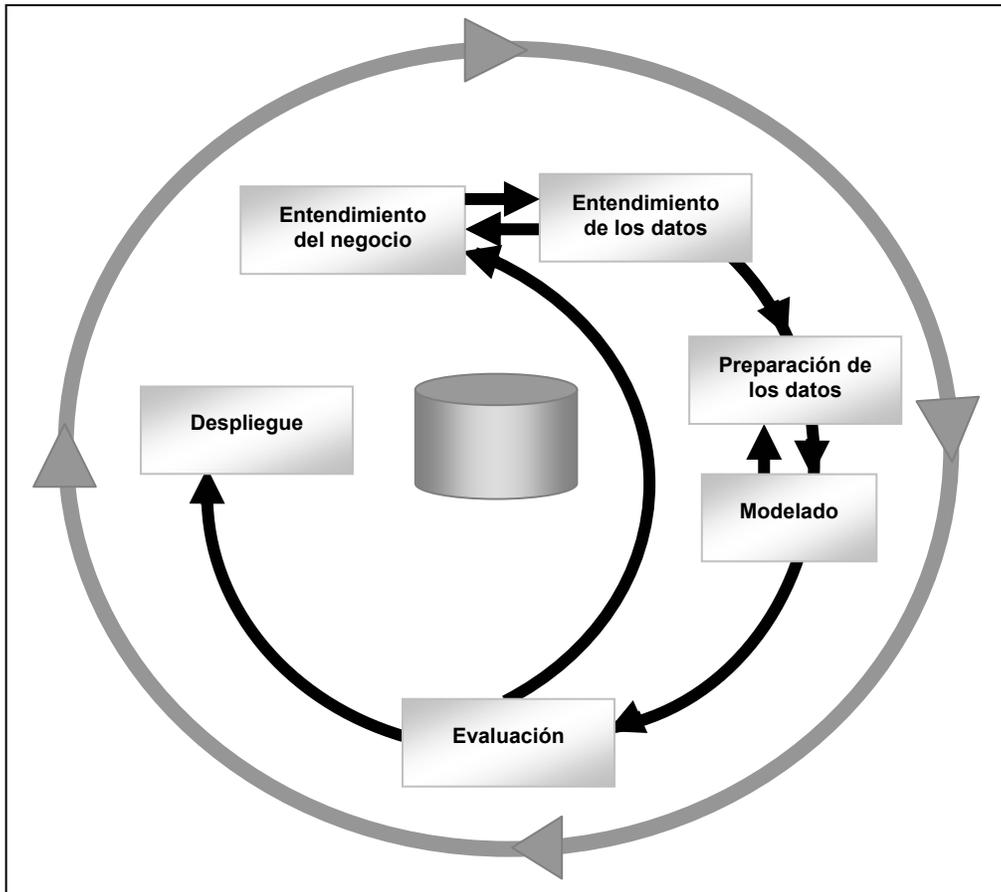
La metodología no implica pasos en cadena, normalmente una fase no ha finalizado cuando inicia la siguiente, es importante subrayar que la Minería de Datos no termina cuando se tienen una solución, se deben retroalimentar los datos, prepararlos nuevamente y volver a generar las fases en repetidas ocasiones, hasta obtener el resultado esperado.

Es pertinente indicar que la organización cuenta con una herramienta (software) de Minería de Datos, ésta es usada para desarrollar la parte del proyecto relacionada con minado de datos, el software se llama "Intelligent Miner" (IM) de la compañía IBM.

Los algoritmos de Minería de Datos disponibles dentro del software IM son los siguientes:

- Clasificación
- Reglas de asociación
- Segmentación
- Predicción
- Patrones secuenciales

- Secuencias similares



Fases de la Minería de Datos

Figura 3.3

También incluye algunas otras técnicas para la detección de diferentes tipos de correlaciones o patrones dentro de la base de datos, y son las siguientes:

- Análisis de componentes principales
- Análisis de factores
- Regresión

El reto no sólo es determinar cuál técnica utilizar sino cómo debe aplicarse, ya que esto resulta complejo si no se tiene un conocimiento profundo de los datos y de los resultados que genera cada algoritmo, pero esto se lleva a cabo en la siguiente sección.

3.2 PROCESO GENERACION DE RECOMENDACIONES

Como se revisó en el capítulo uno, existen muchas propuestas y trabajos desarrollados para la generación de recomendaciones, cada uno con características particulares.

Uno de los sistemas de recomendaciones más parecidos al problema que se está desarrollando es el propuesto por *R. D. Lawrence et. al* (2001), el sistema da como resultado la sugerencia de productos en un supermercado basado en el comportamiento de compras previas. Este proceso será utilizado como base para el desarrollo de este trabajo, por razones de compatibilidad tanto de situación como de software.

Proponen asociaciones basadas en la taxonomía de los productos, la cual se encuentra almacenada en la base de datos como clase y subclase, de tal forma que éstas se usan como característica de productos así como también en los perfiles. Una persona muestra interés en una característica al comprar productos que tienen clase y subclase.

Los productos para recomendar se determinan calculando una medida de distancia entre vectores que representan preferencias personales y vectores que representan productos.

Se usa filtrado colaborativo, de tal manera que se agrupa el consumo de personas con gustos similares, de aquí surgen dos fuentes de información:

1. Reglas de asociación de minería de datos. Se aplican a los datos de las compras realizadas por clientes, para derivar relaciones entre clases y subclases de productos.

2. Segmentación de clientes. Agrupa clientes con intereses similares, cada cliente se asigna a un solo segmento.

El sistema recomendador usa filtrado contextual con ideas de filtrado colaborativo, utiliza ambos para refinar el modelo contextual y para hacer recomendaciones de intereses compartidos con segmentos de clientes.

Como medio de comunicación con el cliente se utilizan PDAs (Personal Digital Assistant), y genera recomendaciones semanales para cada uno, almacenándolas en la memoria de las PDA, de esta manera los clientes tienen en sus manos la información más nueva.

Las recomendaciones para clientes con objetivos similares son unificadas en una lista de productos populares dentro del segmento asignado. La lista de recomendaciones pasa por una "Ingeniería de Acoplamiento", la cual ordena esta lista de acuerdo a la atracción esperada para cada cliente. El acoplamiento usa relaciones de productos o listas de recomendaciones personalizadas, el cual toma de 10 a 20 productos con la calificación más alta, además la lista contiene productos que no han sido comprados.

Tomando el trabajo de Lawrence como base, se desarrolla un proceso similar, que comprende desde la generación de información fuente hasta la lista de recomendaciones para cada cliente.

Con la información de las compras de los clientes se producen los dos vectores de características que soportan el proceso:

- (1) **El de productos afines.**
- (2) **El de hábitos de consumo.**

Que se calculan con base en el filtrado colaborativo, posteriormente, la obtención de una lista de recomendaciones para cada cliente es un procedimiento relativamente sencillo; sin embargo, la lista resultante puede ser refinada utilizando otros elementos para lograrlo.

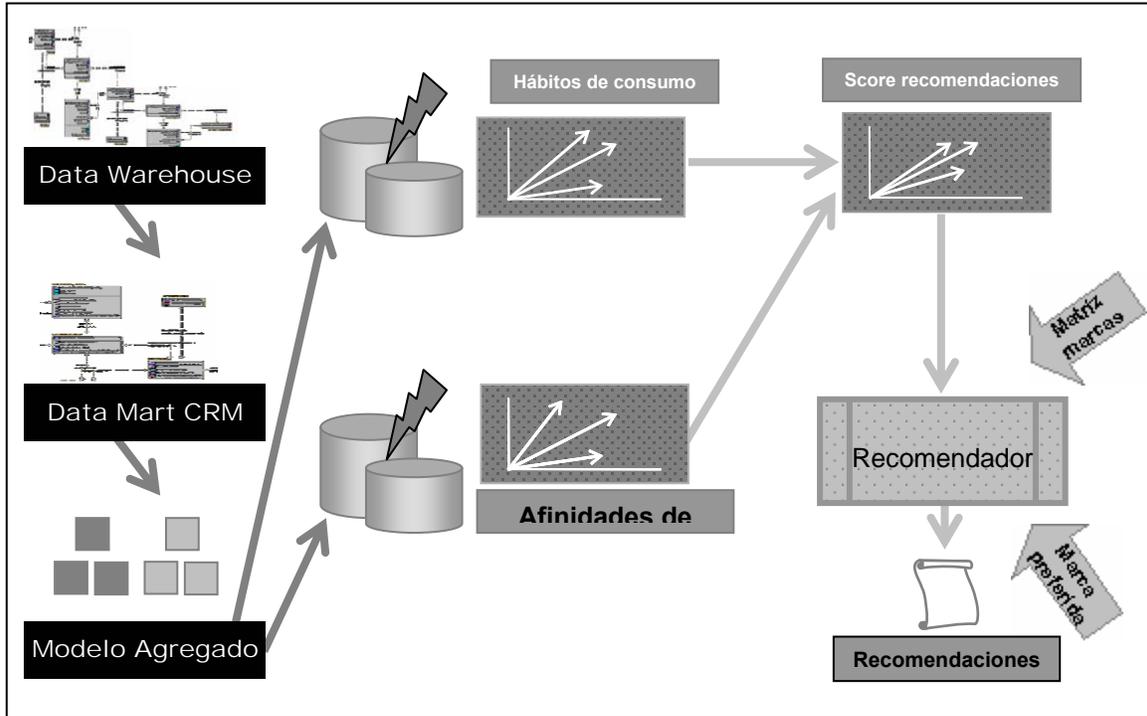
Para tal efecto, los productos recomendados pueden estar acompañados de una marca. Para determinar que marca sugerir, puede construirse un espacio de preferencias de marcas, que al ser combinado con la matriz de marcas, permite recomendar marcas inclusive en oportunidades de cross-selling (venta cruzada).

El proceso para generar recomendaciones personalizadas está formado por los siguientes elementos:

- **Información:** Datos que son la fuente de información, tanto para conocer los hábitos de consumo de cada cliente, como para encontrar las afinidades entre productos.
- **Modelos:** Que determinan los hábitos de consumo, así como las afinidades entre productos.
- **Recomendaciones:** A partir de los hábitos de consumo y de las relaciones entre productos se genera una lista de productos afines con una calificación, esta lista es refinada y finalmente si es posible se agrega la marca preferida o afín.

La figura 3.4 muestra el proceso seguido para la generación de recomendaciones, desde la extracción y generación de datos, el desarrollo del proceso de afinidades o relaciones de productos, el proceso de generación de hábitos de consumo de los clientes, la asignación de peso de las recomendaciones hasta la generación de recomendaciones específicas.

El sistema diseñado utiliza el filtrado contextual como parte central, además del filtrado colaborativo y otros elementos de información particulares del negocio que contribuyen de manera significativa en el refinamiento del resultado que produce el modelo.



Proceso Generación de Recomendaciones Personalizadas

Figura 3.4

3.3 ANÁLISIS DE INFORMACIÓN

Las reglas del negocio dan la dirección sobre que variables utilizar y el significado de las mismas, a continuación se describen las características propias del negocio relevantes para el problema y la información relacionada con la que se cuenta, se muestran las consideraciones tomadas en cuenta durante el desarrollo del proceso, tales como los periodos de tiempo, los clientes o la taxonomía de productos.

- **Periodo de análisis**

En un entorno económico relativamente estable como el que hemos tenido los últimos años en México, el no realizar ajustes por inflación en un periodo corto de doce meses, representa grandes ventajas al favorecer en la simplicidad del modelo (en cuanto a datos y tiempo de procesamiento), y sin incurrir en anomalías graves con las variables asociadas al consumo. No se debe perder de vista que para algunos productos los ciclos de compra tienen un fuerte vínculo con aspectos estacionales, como los adornos de navidad o los trajes de baño.

Se eligió un horizonte de tiempo razonable para la detección de hábitos de consumo, de doce meses, que va de Enero de 2005 a Diciembre de 2005, este período contempla todas las épocas del año, por lo cual engloba todo tipo de compras de características temporales.

- **Frecuencia de recomendaciones**

El plan es crear una lista de recomendaciones para cada cliente mensualmente, y para cálculos futuros tanto para los hábitos de consumo como para el minado de datos, se considera utilizar periodos de 12 meses móviles con el fin de integrar siempre todas las épocas del año.

- **Selección de clientes**

El universo de clientes para este análisis es el formado por los individuos que poseen una tarjeta de crédito de la tienda departamental, tanto titulares como adicionales, y que además hayan realizado al menos una transacción de compra en el período de análisis. El número de clientes que cumplen esta condición es 332,470.

- **Grado de personalización**

La información transaccional fue consolidada a nivel tarjeta de crédito, buscando que los perfiles de hábitos de consumo contruidos tengan un suficiente nivel de individualidad, evitando el efecto de integrar el gasto de un grupo en uno solo, que puede producirse al consolidar la información a un nivel más alto como es el caso de la línea de crédito, que contempla a una tarjeta titular y varias adicionales. Aún así, pueden existen casos en que una tarjeta es utilizada para hacer compras de más de una persona, puede ser de una pareja o de una familia, en todo caso las recomendaciones serán dirigidas al titular.

- **Unificación de departamentos**

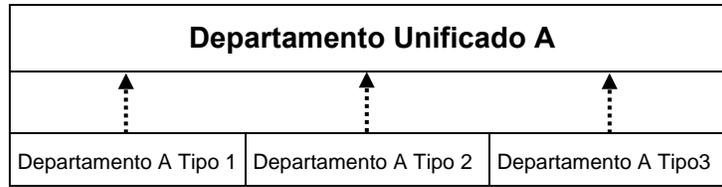
Debido a cuestiones administrativas de la compra de mercancías, el catálogo de departamentos permite la diferenciación de la forma en la que son adquiridos los productos a nivel departamento, considerando diferentes opciones; sin embargo, esta separación es irrelevante en términos de la detección de hábitos de consumo. Por este motivo, se realizó la consolidación de dichos casos hacia un departamento unificado (Tabla 3.1). Esto permite integrar todas las compras realizadas mediante una misma tarjeta sin importar el manejo administrativo de los productos.

La unificación de departamentos reduce la cardinalidad del conjunto de 241 a 126 departamentos, con esta integración se busca mejorar la posibilidad de capturar los hábitos de consumo.

- **Unificación de estilos**

Derivado también de cuestiones administrativas se presenta el caso de que un mismo estilo puede existir en múltiples ocasiones. De manera equivalente con el departamento se realiza una

integración automática, que refleja una disminución significativa en el volumen de estilos registrados. La unificación de estilos reduce la cardinalidad del conjunto de 1'177,051a 823,922 estilos, y esto se refleja directamente en los tiempos de ejecución de los procesos.



Departamentos Unificados

Tabla 3.1

- **Taxonomía de productos**

La agrupación de mercancías que se tiene hoy en día considera los siguientes niveles (Tabla 3.2).

5	Departamento
4	Clase
3	Subclase
2	Estilo
1	Producto

Taxonomía de productos a nivel departamento

Tabla 3.2

Cuando se está en la búsqueda de hábitos de consumo y patrones de comportamiento resulta conveniente agrupar los productos con un enfoque diferente. Por esta razón se incluye una rama alternativa que parte del departamento y que considera en los niveles superiores Categoría y Súper-categoría.

Las estructuras integradas tienen una cardinalidad de 30 y 16 elementos respectivamente. Este enfoque no interfiere con la estructura actual, pero sí favorece de manera importante el análisis de información transaccional con un punto de vista orientado al descubrimiento y asociación o asignación de posibles perfiles.

- **Nivel de recomendación**

El nivel de detalle de la recomendación tiene una relación directa con la construcción de los hábitos de compra, las asociaciones y su manejo dentro de la base de datos, esto hace que la cantidad de miembros de esta dimensión no sea excesiva. Por otra parte, el catálogo de productos al nivel más bajo presenta características tan detalladas como la talla y el color. La combinación de ambos

factores impulsa a elegir un nivel intermedio de la taxonomía para la generación de las recomendaciones prudentes.

Se decidió tomar el nivel jerárquico clase, que a pesar de poseer una cardinalidad de 1,815 elementos, solo se tiene evidencia de 493 clases adquiridas durante el periodo de análisis, ya con la utilización de departamento unificado.

- **Consolidación temporal**

Un elemento que sin duda tiene un fuerte efecto en la posibilidad de detección de afinidades entre productos a través de la Minería de Datos, tiene que ver con el periodo de tiempo para incluir las transacciones realizadas. Por la propia naturaleza de una tienda departamental, en términos generales las transacciones puntuales no arrojan ninguna evidencia de vínculos fuertes entre productos u oportunidades de cross-selling (venta cruzada). Además de la existencia de productos con patrones de consumo estacionales, también pueden llegar a detectarse ciertas relaciones entre productos que en términos de negocio no son viables, como el vínculo entre un abrigo y un traje de baño.

Por esta razón se agregaron las transacciones de un período de tiempo, con el fin de aumentar la posibilidad de detectar reglas de asociación más atractivas. La decisión tomada para acumular las transacciones contempla periodos trimestrales.

Es decir que en lugar de tener el consumo del cliente para cada uno de los tickets, se acumulan todos los que fueron consumidos en un período de tres meses, se maneja como si el cliente consumiera un “ticket trimestral”, por lo cual un cliente puede tener de 1 a 4 en un período de un año.

- **Diversidad de consumo**

Para incrementar la detección de asociaciones de productos en términos de ventas cruzadas, se recomienda hacer un filtrado de transacciones utilizando sólo aquellas que pertenecen a clientes que realizan transacciones en departamentos o clases diversas. Esto con el fin de quedarse sólo con las transacciones que reflejen diversidad en las compras. Clientes con compras de un solo producto son eliminadas de las asociaciones, por ejemplo un cliente que sólo ha comprado una tarjeta de teléfono no aporta información al filtrado colaborativo. Con lo anterior se impone una restricción que favorece la consolidación de compras con una cartera de productos más diversa.

Por otra parte se encuentra el porcentaje máximo de consumo en cualquier categoría. Por ejemplo, un cliente con un consumo de 94% en la categoría A y un 6% de consumo repartido en las categorías B, C, D, E, F y G; en este caso la diversidad del cliente es secundaria.

Por lo cual se eligió tomar las transacciones de los clientes con compras en 5 distintas categorías, y con 50% para el máximo de consumo en cualquiera de ellas.

- **Valor de la mercancía**

Para evitar tener en consideración un alto volumen de asociaciones de productos potencialmente irrelevantes, se siguió un proceso mediante el cuál se descartaron aquellos productos que a pesar de su alto volumen transaccional, no contribuyen a realizar recomendaciones importantes para los clientes, y además que no representan en términos económicos un impacto positivo para la tienda. Dentro del departamento de Gourmet los dulces, chocolates y galletas son muy populares, muchos clientes compran alguna vez éste tipo de productos básicamente porque son de bajo costo, esto provoca que las galletas estén asociadas con todo tipo de cosas, razón por la cual se descartaron aquellos productos que al ser adquiridos no aportan un valor económico importante de manera individual, tomando en cuenta sólo aquellos productos con un valor monetario mínimo de \$100.

3.4 ALGORITMO DE REGLAS DE ASOCIACION

Las reglas de asociación permiten descubrir las combinaciones de productos que los clientes compran y las relaciones que existen en todos los niveles en la jerarquía de productos. Esto incluye reglas específicas que digan que productos concretos se relaciona con otro, por ejemplo Chateau Hursley (1999) revela que un vino tinto particular se compra en combinación con un queso específico. Otro caso es que los vinos tintos alemanes se compran con queso francés, o para un nivel más general simplemente se dice que vino y queso se compran juntos.

Las reglas de asociación expresan patrones de comportamiento entre los datos en función de la aparición conjunta de valores de dos o más atributos. La característica de éstas es que manejan atributos nominales, a diferencia de otros estudios de correlaciones numéricas.

Las reglas de asociación en Minería de Datos expresan combinaciones de los atributos que suceden más frecuentemente, las cuales surgieron inicialmente para enfrentar el “Market Basket Analysis” o análisis de las Cestas de Compra de los comercios.

Una regla de asociación tiene la forma:

Si α entonces β , donde α y β son dos conjuntos de artículos disjuntos.

Y se puede expresar como $\alpha \Rightarrow \beta$.

En donde el lado izquierdo implica el lado derecho. En el contexto de “Market Basket Analysis” se dice que si el producto α se compra, entonces implica que el producto β será también comprado al mismo tiempo.

El conjunto α es el predecesor de la regla, comúnmente llamado cuerpo de la regla (Body), que puede contener uno o múltiples artículos.

El conjunto β es el sucesor de la regla, comúnmente llamado cabeza de la regla (Head), que contiene un artículo.

En la técnica de Reglas de Asociación usualmente se tienen dos medidas para conocer la calidad de la regla, pero el software utilizado agrega la confianza de la regla.

- **Soporte.** Es el número de instancias que la regla preside correctamente (también se usa el porcentaje). Es decir mide la fracción de cestas (en este caso tickets trimestrales) que contienen ambos el lado izquierdo y el lado derecho de la regla. Si el producto α y el producto β son encontrados en el 10% de las cestas, entonces el soporte es 0.1.
- **Confianza.** Mide el porcentaje de veces que la regla se cumple cuando se puede aplicar. En otras palabras, mide la fracción de la canasta (ticket trimestral) en el lado izquierdo que también contiene el producto en el lado derecho de la regla. Si el producto β está presente en 50% de las canastas y contienen el producto α , entonces la confianza es 0.5. Si se sabe que el producto α está en una cesta particular, entonces el producto β también es encontrado en la misma cesta el 50% de las veces.
- **Lift o Confianza de la regla.** Esta medida se encuentra en la herramienta de minería usada (Intelligent Miner). Es la confianza dividida entre el número de casos cubiertos por la parte derecha de la regla. Es decir la fracción de la confianza de la regla para la confianza esperada de encontrar el lado derecho en cualquier cesta (ticket trimestral). Por ejemplo, si el producto β fue encontrado en sólo 5% de todas las cestas, entonces el Lift para la regla tiene el valor 10. En 5% de los casos donde α y β están en la misma cesta, entonces la asociación ocurre 10 veces más frecuente.

El algoritmo A priori es el más simple y común para calcular Reglas de Asociación con Minería de Datos, se basa en la búsqueda de los conjuntos de artículos (items) con determinado soporte. El funcionamiento de este algoritmo se basa en la búsqueda de los conjuntos de artículos con determinado soporte.

En primer lugar se construyen los conjuntos formados por un solo artículo que supera el soporte mínimo. Este conjunto de conjuntos se utiliza para construir el conjunto de dos artículos, y así sucesivamente hasta que se llegue a un tamaño en el cual no existan conjuntos de artículos con la cobertura requerida.

La generación de reglas de asociación se divide en dos fases: la extracción de conjuntos de artículos que cumplan con el soporte requerido y la generación de reglas a partir de estos conjuntos.

La selección de candidatos de este algoritmo consiste en formar, dado un conjunto de artículos de tamaño i , los posibles candidatos de tamaño $i+1$. Si se dispone de un grupo Z formado por conjuntos de artículos de tamaño i , para la selección de candidatos, se toma cada pareja $\{x, y\}$ que formen un conjunto de artículos de tamaño $i+1$, donde x y y son conjuntos de artículos de Z .

La siguiente fase consiste en la creación de reglas a partir de los conjuntos de artículos frecuentes. Si sólo buscamos reglas de asociación con un artículo en la parte derecha, el proceso es sencillo: de un conjunto de artículos de tamaño i , se crean i reglas colocando siempre un único artículo diferente en la parte derecha.

A continuación se muestra el algoritmo **A priori** en la figura 3.5.

```

ALGORITMO A priori (D: datos, MinC: cobertura mínima)
i = 0
Rellena_Item( $C_i$ ) //Incluyen en  $C_0$  todos los items de tamaño 1
MIENTRAS  $C_i \neq \emptyset$ 
  PARA CADA x = elemento de  $C_i$ 
    SI Cobertura (x)  $\geq$  MinC ENTONCES  $L_i = L_i \cup x$ 
  FINPARA
   $C_{i+1} =$  Selecciona_Candidatos ( $L_i$ )
  i = i+1;
FIN MIENTRAS
RETORNA C
FIN ALGORITMO
    
```

Algoritmo A priori

Figura 3.5

En este trabajo se usan reglas de asociación con el algoritmo A priori, que se encuentra en el software Intelligent Miner.

En su trabajo, Lawrence (2001) opera una tabla o entidad de entrada con dos elementos: (1) Id del Cliente y (2) Id de la Clase, cada producto se representa por medio de su clase y departamento en que el cliente ha comprado en el periodo. Se calculan sólo asociaciones con un artículo en el cuerpo de la regla y uno en la cabeza.

Como ya se ha mencionado previamente para ejecutar el algoritmo de asociación de productos se ha utilizado filtrado en dos sentidos:

- El filtrado contextual.
- El filtrado colaborativo.

Los cuales se describen a continuación.

- **Filtrado contextual**

Un sistema de filtrado contextual produce recomendaciones basado en la similitud de lo que un individuo dado ha preferido en el pasado. Típicamente, tanto los artículos como los perfiles son representados como vectores de un espacio de características, y su similitud es calculada mediante una medida de distancia estandarizada.

Para utilizar el filtrado contextual es necesario definir el espacio de características, en el cual serán situados tanto los hábitos de consumo individuales, como las afinidades entre productos.

El detalle de las transacciones realizadas por los clientes está referido al nivel producto; sin embargo, queda fuera de alcance considerar la posibilidad de construir el componente vectorial de hábitos de consumo a dicho nivel, debido a su limitada capacidad para la detección de hábitos y a la extrema complejidad computacional resultante.

Se decidió que el nivel jerárquico departamento-clase es más que suficiente, ya que basta con saber que un cliente compró una blusa en el departamento damas y no tomar el detalle que fue una blusa azul con flores blancas, con botones al frente, talla 34. Por este motivo se decidió proceder con un grado de consolidación al nivel jerárquico departamento-clase.

Se considera que una persona muestra interés en una característica específica al adquirir productos pertenecientes a su clase correspondiente. Los productos a recomendar pueden determinarse calculando una medida de distancia entre el vector que representa las preferencias personales y los vectores que representan los productos.

Este enfoque por sí solo presenta un problema llamado sobre-especialización, tratado por Balabanovic y Shoham (1997), ya que no proporciona bases rigurosas y suficientes para tratar de inducir a los compradores hacia nuevas clases de productos a partir las clases que han consumido. Para resolver este problema pueden incorporarse características que vinculen a productos relacionados, lo cual muestra la flexibilidad de este enfoque, pero requiere necesariamente de intervención humana para detectar, diseñar e incorporar las características de vínculo. En un ambiente de negocios dinámico, diverso y con un alto volumen transaccional, lo anterior no es posible.

- **Filtrado Colaborativo**

Este enfoque busca evitar el problema del diseño de características de vínculo recomendando artículos que otros individuos, que son similares al cliente, han adquirido anteriormente. Sistemas de filtrado colaborativo como “Ringo”, propuesto por Shardanand y Maes (1995) o “GroupLens” por Konstan (1997) funcionan recolectando calificaciones de usuario explícitas respecto a los artículos en cuestión. Los clientes son comparados apoyándose en la similitud de las calificaciones que otorgan y entonces se recomiendan artículos que son favorecidos por otros individuos con intereses similares.

Para construir el componente de afinidades de productos en el espacio de características, Agrawal y Srikant (1994) aplican técnicas de Minería de Datos que permiten detectar relaciones entre clases y departamentos.

Las relaciones están basadas en compras hechas por los clientes, por lo que se espera detectar vínculos desconocidos y diferentes a los que pueden ser expuestos de manera natural por la propia

taxonomía. Por sí mismo, el filtrado contextual no incorpora información acerca de la popularidad relativa de los productos mostrada por otros clientes.

- **Detección de reglas**

Con una amplia gama de productos y clientes comprándolos, es difícil decidir qué productos recomendar a cada cliente. Una campaña para promover nuevos productos y servicios es del tipo cross-selling (ventas cruzadas), mientras que una para promover productos y servicios más rentables o de mayor margen es del tipo up-selling (ventas adicionales). En este contexto la mejor recomendación significa productos relevantes o interesantes para el cliente, para que al cliente le resulte individualizada, y una inusual o interesante le resulte emocionante. Para lograr esto es necesario encontrar aquellos productos afines o de interés para el cliente.

La determinación de las afinidades de productos para la generación de recomendaciones puede provenir de tres orígenes:

- La relación jerárquica de los productos: sin evidencia histórica o naturaleza de la mercancía.
- La detección de reglas de asociación realizada mediante algoritmos de Minería de Datos: evidencia histórica o impulso real.
- El diseño predeterminado de vínculos: evidencia empírica o impulso voluntario.

Dependiendo del tipo de oportunidades de venta que se desean impulsar puede utilizarse un conjunto de reglas para poner en juego estos diferentes elementos.

Como ya se mencionó, para este trabajo se toma de la taxonomía el Departamento y la Clase, haciendo las recomendaciones ligadas con Departamento y Clase.

A partir de las reglas de asociaciones se puede determinar la relación o vínculo entre productos, pero también existen otros tipos de posibles conexiones que determinan enfoques para generar tipo de relaciones entre productos.

Las reglas entre clases de un mismo departamento se designan con el nombre de "Intradepartamentales".

Las reglas entre clases de departamentos distintos como "Interadepartamentales".

Un ejemplo de este tipo de reglas es la relación que hay entre tipos de vinos del departamento de Vinos y Licores para el primer caso, y la relación que existe entre vinos y queso de los departamentos de Vinos y Licores y Gourmet (Figura 3.6).

Los tipos de vínculos entre clases y departamentos se pueden clasificar en 6 tipos de acuerdo al tipo de relación, nivel jerárquico, enfoque y el origen de la misma, es decir si fue derivada del algoritmo de Minería de Datos o no. Esta clasificación más detallada se resume en la tabla 3.3.

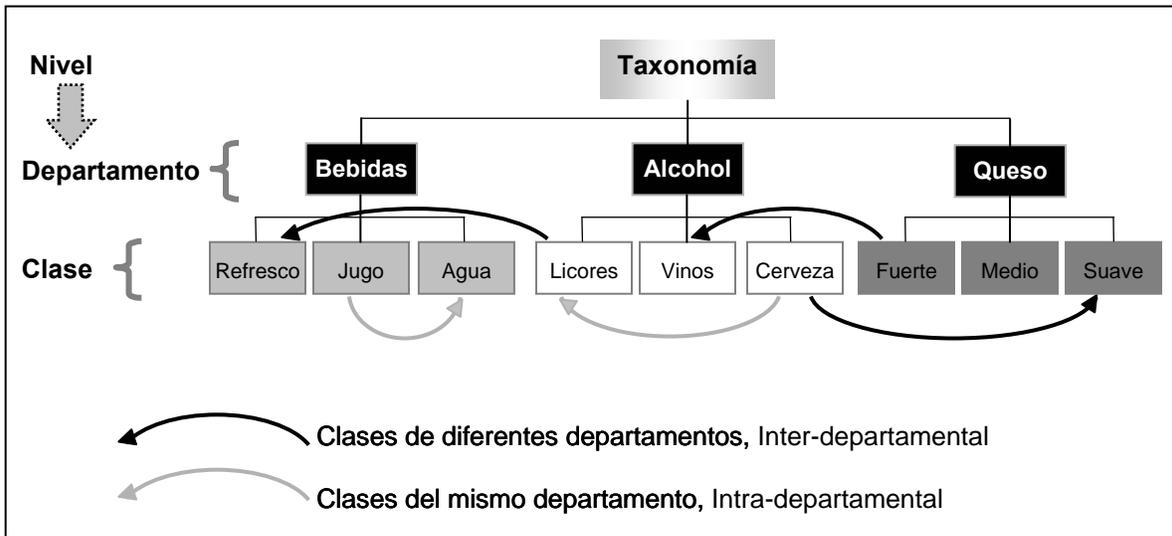


Figura 3.6
Vínculos entre productos

	Tipo de vínculo	Nivel	Enfoque	Origen
A	Vínculo fuerte	Clase	Inter-departamental	Con evidencia. Reglas de Asociación.
B	Vínculo fuerte	Clase	Intra-departamental	Con evidencia. Reglas de Asociación.
C	Vínculo taxonómico	Clase	Intra-departamental	Sin evidencia. Por la jerarquía propia de lo productos.
D	Vínculo débil	Depto.	Inter-departamental	Con evidencia. Reglas de Asociación.
E	Auto-vínculo	Clase	Relación auto relacional	Sin evidencia. Por la jerarquía propia de lo productos.
F	Vínculo diseñado	Clase	Relación exógena	Diseño predefinido

Clasificación de vínculos entre productos

Tabla 3.3

Los vínculos de los tipos A, B y D se obtienen con las reglas de asociación de minería de datos. Los vínculos de los tipos C y E se obtienen a partir de la estructura jerárquica de productos, las clases de un departamento tienen una relación, por el solo hecho de pertenecer al mismo departamento.

El vínculo del tipo F es un elemento externo, en este caso existe una estrategia para promover productos. Se diseña previamente y con el propósito de promover una clase.

Siguiendo el ejemplo de la figura 3.6 e integrándolo en los tipos de relaciones, se muestran ejemplos de cada uno de los tipos de vínculo en tabla 3.4.

<p>A. Vínculo fuerte ínter</p> <p>Licores → refresco Queso fuerte → Vinos Cerveza → Queso suave</p>	<p>B. Vínculo fuerte íntra</p> <p>Jugo → Agua Cerveza → Licores</p>	<p>C. Vínculo taxonómico</p> <p>Refresco ↔ Jugo Refresco ↔ Agua Jugo ↔ Agua Etc.</p>
<p>D. Vínculo débil</p> <p>Vinos → Agua Vinos → Jugo Vinos → Refresco Etc.</p>	<p>E. Auto-vínculo</p> <p>Refresco → Refresco Vinos → Vinos Queso fuerte → Queso fuerte Etc.</p>	<p>F. Vínculo diseñado</p> <p>Refresco → Licores Queso suave → Refresco</p>

Ejemplo de tipos de relaciones

Tabla 3.4

Para la concepción de recomendaciones en este trabajo se utilizan sólo vínculos provenientes de procedimientos automáticos, es decir, vía relación jerárquica o vía minería de datos. De esta forma, se cubren los primeros cinco tipos de vínculos.

El algoritmo de asociaciones necesita datos de entrada que son la base para la generación de reglas de asociación. Para esto se han elaborado varios conjuntos de datos con las características de las que ya se ha hablado, así, las reglas de asociación se generan bajo las siguientes condiciones:

- El cliente se identifica por su tarjeta.
- Las compras del cliente se acumulan trimestralmente, agregando todas sus compras a una compra trimestral.
- Los productos comprados se catalogan a nivel departamento-clase.
- Se toman aquellos productos que tienen valor comercial mayor a \$100.
- Se toman los clientes en cuyas transacciones existe una diversidad de menos de 50% en un departamento-clase, con el fin de que no existan transacciones de un solo producto.

La información que alimenta el algoritmo de Reglas de Asociación está formado por cliente y producto, es decir que cada clase comprada en el trimestre por el cliente se representa con un renglón en la tabla. Por ejemplo si un cliente compró vino tinto en cada trimestre se generan 4 renglones diferentes para el cliente con la clase vino tinto.

En cuanto a la longitud de las reglas de asociación, sólo se toman reglas de un producto, es decir uno a uno.

Después de realizar varios ejercicios en el entendimiento, preparación de datos y modelado, se ha llegado al entendimiento de que el departamento de Perfumería está relacionado con todos los productos y aparece frecuentemente, esto ha ayudado a entender el negocio un poco más fortaleciendo el hecho de que es un departamento de impulso, lo cual lleva a probar escenarios con y sin incluir el departamento de Perfumería en nuevos ejercicios de Asociaciones con el software, esto siguiendo la metodología genérica.

Después de generar diversos grupos de datos y de hacer repetidas pruebas con el algoritmo, se llega a un resultado satisfactorio. La tabla 3.5 muestra el volumen de la información que se procesó en el algoritmo, para generar las reglas de asociación finales durante el minado de datos.

	Ocurrencias
Tarjetas participantes	332,470
Detalle de transacciones	6'951,258
Departamentos	241
Clases	1,815
Subclases	9,091
Estilos	1'177,051
Marcas	7,804

Volumen de Información para el Minado de Datos

Tabla 3.5

Los detalles y parámetros que dieron origen a las reglas de asociación finales se encuentran resumidos en la tabla 3.6. Los escenarios se refieren a las características que cumplen las reglas de asociaciones finales:

Evento Compra Trimestral Clase: Incluye todas las relaciones encontradas al nivel jerárquico Departamento-Clase.

Evento Compra Trimestral Clase sin Perfumería: Incluye todas la relaciones encontradas al nivel jerárquico Departamento-Clase, eliminando todos los productos pertenecientes al departamento Perfumería.

Evento Compra Trimestral Departamento: Incluye todas las relaciones encontradas al nivel jerárquico Departamento.

Evento Compra Trimestral Departamento sin Perfumería: Incluye todas las relaciones encontradas al nivel jerárquico Departamento, eliminando el departamento Perfumería.

Herramienta de minería	Intelligent Miner IBM.
Algoritmo	Reglas de asociación – Market Basket Analysis Algoritmo de asociaciones “A priori”
Base de minado	GRP V1.0
Escenarios	AR EVENTO COMPRA TRIMESTRAL CLASE AR EVENTO COMPRA TRIMESTRAL CLASE SIN PERF AR EVENTO COMPRA TRIMESTRAL DEPTO AR EVENTO COMPRA TRIMESTRAL DEPTO SIN PERF
Parámetros	Soporte: 0.25% Confianza: 25% Longitud máxima de la regla: 2

Detalles y Parámetros para las Reglas de Asociación Finales

Tabla 3.6

A continuación se presentan algunos ejemplos de las relaciones encontradas con las reglas de asociación (Tabla 3.7). Los productos se muestran con la notación Departamento//Clase, el primero se presenta con mayúsculas y la clase con minúsculas, la separación entre ambos es una doble diagonal “//”, en la tabla se muestra el soporte, la confianza y el lift, datos de los cuales se habla al inicio de esta sección.

Body	Header	Soporte	Confianza	Lift
[PERFUMERIA // Tratamiento P/Dama]	[PERFUMERIA // Maquillaje]	0.05163	0.4890	3.5505
[SPORT CABALLERO // Camisa]	[SPORT CABALLERO // Pantalón]	0.02419	0.3174	3.4941
[DISCOS // DVD Y VHS]	[DISCOS // Disco compacto]	0.02324	0.3340	4.3297
[JUGUETERIA // Juego]	[JUGUETERIA // Juguete Niño]	0.02101	0.3941	4.4081
[PERFUMERIA // Fragancia Femenina]	[PERFUMERIA // Maquillaje]	0.02087	0.2947	2.1397
[DAMA // Pantalón]	[DAMA // Blusa]	0.02070	0.4491	4.5551
[SPORT CABALLERO // Playera]	[SPORT CABALLERO // Pantalón]	0.02015	0.2659	2.9272

Ejemplo relaciones encontradas entre productos

Tabla 3.7

Body	Header	Soporte	Confianza	Lift
[JUGUETERIA // Juguete Niña]	[JUGUETERIA // Juguete Niño]	0.01760	0.3613	4.0412
[DAMA // Saco]	[DAMA // Pantalón]	0.01707	0.4783	10.3784
[JUNIOR Y MISSES // Pantalón]	[JUNIOR Y MISSES // Blusa]	0.01524	0.3554	7.2073
[DAMA // Saco]	[DAMA // Blusa]	0.01409	0.3949	4.0054
[BONETERIA Y ACCESORIOS // Corbata]	[CAMISA Y SWEATER//Camisa vestir color]	0.01409	0.3004	5.4683
[PERFUMERIA // Fragancia Masculina]	[PERFUMERIA // Fragancia Femenina]	0.01354	0.2850	4.0249
[JUGUETERIA // Juguete Niña]	[JUGUETERIA // Juego]	0.01276	0.2619	4.9119
[JUGUETERIA // Muñeca]	[JUGUETERIA // Juguete Niña]	0.01217	0.2984	6.1257
[ACCESORIOS DAMA // Media]	[PERFUMERIA // Maquillaje]	0.01130	0.2999	2.1775
[ROPA DEPORTIVA // Ropa Caballero]	[ROPA DEPORTIVA // Ropa Dama]	0.01117	0.4527	8.5074
[VINOS Y LICORES // Aguardiente]	[ALIMENTOS GRAV. VTA // Alimentos]	0.01089	0.2661	1.6986
[JUGUETERIA // Video Juego]	[JUGUETERIA // Juguete Niño]	0.00959	0.2876	3.2169
[JUNIOR Y MISSES // Falda]	[DAMA // Blusa]	0.00931	0.3707	3.7599
[DAMA // Falda]	[DAMA // Blusa]	0.00928	0.4946	5.0166
[JOYERIA Y RELOJES // Fantasía]	[PERFUMERIA // Maquillaje]	0.00901	0.2618	1.9008
[NIÑAS // Pantalón]	[NIÑAS // Deportivo]	0.00812	0.3632	11.7028
[JUNIOR Y MISSES // Top]	[JUNIOR Y MISSES // Blusa]	0.00809	0.3082	6.2501
[BEBES // Playera y Sudadera]	[BEBES // Pantalón]	0.00793	0.3203	9.6175

Ejemplo relaciones encontradas entre productos

Tabla 3.7 (continuación)

3.5 CONSTRUCCIÓN DE PERFILES DE CONSUMO

En esta sección se analiza la manera en que los clientes realizan su gasto, es decir sus hábitos o perfil de consumo. Lawrence (2001) hace una segmentación de clientes para recomendar listas a los clientes pertenecientes a un su segmento, para el problema que aquí se desarrolla no se hace segmentación, se trabaja con los perfiles de cada uno de los clientes de manera individual, así cada cliente tiene un perfil particular y se asocia una lista de recomendaciones única.

3.5.1 Consumo de productos

Una vez que se generan los vectores de hábitos de compra o consumo para cada cliente, se acoplan las compras del cliente con las reglas de asociación encontradas, es decir para cada

Departamento-Clase que el cliente haya adquirido se asocia con el cuerpo (Body) que corresponde, en caso de que sea posible, y se toma la cabeza (Header) de la Regla, de esta manera se genera una lista de recomendaciones inicial, cuyos elementos están total y únicamente ligados a cada persona, ya que dependen de las compras específicas de cada cliente y no de las compras conjuntas de un grupo de clientes pertenecientes a un segmento.

Los hábitos de compra de cada cliente se plasman en vectores de consumo, que contienen artículos que el cliente ha adquirido, estos se construyen con los datos siguientes:

- Tarjeta
- Departamento unificado
- Clase
- Gasto en la clase
- Precio de lista
- Costo
- Porcentaje de gasto en la clase

Cada uno de estos vectores (uno por cada cliente) se cruzan con las reglas de asociación, para esto se utiliza una heurística que conjunta ambos tipos de filtrado (colaborativo y contextual) con los tipos de vínculos (íntra e Inter Departamentales), la heurística es una de las partes más relevantes en la generación de recomendaciones, ya que agrega un componente de valoración o calificación junto con las asociaciones de productos encontradas a través de la Minería de Datos.

3.5.2 Marcas

Por otra parte se tiene la marca favorita del cliente, éste es un nuevo concepto que se introduce en este trabajo. No es simple determinar la marca favorita de un cliente, por ejemplo un cliente compra una mezcla de productos que no necesariamente tienen la misma marca, un caso claro es el de una marca de galletas que no tiene una relación alguna con una marca de camisas, por lo cual la definición de marca favorita se delimita sólo con artículos de vestir

La compañía cuenta con una matriz de marcas, que clasifica a las marcas de los departamentos de vestir en dos direcciones, precio y moda, dicha clasificación es para cada Marca asociada a un Departamento.

- **Rango de precio:** Los valores posibles van de lo menos costoso a lo más costoso de la siguiente manera. Bajo, Medio, Alto, Boutiques Bajo, Boutiques Medio y Boutiques Alto.
- **Rango de moda:** Los valores posibles van de lo más conservador a lo menos conservador y con una clasificación por tema de la siguiente manera. Más Clásico, Menos Clásico, Menos Contemporáneo, Más contemporáneo y Temático.

Esta matriz se diseñó sólo para los departamentos unificados pertenecientes a productos de vestir, que son los únicos que pueden clasificarse en un rango de moda, por ejemplo un refrigerador y un queso quedan fuera de ésta clasificación.

Cada Marca perteneciente a un Departamento Unificado se clasifica en una celda, que la define en un rango de precio y moda, así para cada departamento existen 30 celdas que pueden contener más de una marca, y las marcas dentro de cada celda son compatibles.

Para asignar la marca favorita del cliente, se toman los departamentos permitidos y se obtiene aquella donde se ha gastado más.

Así, durante la generación de la lista de recomendaciones, si es posible se agrega la marca favorita o compatible del cliente. En la sección siguiente se habla sobre la heurística que genera las calificaciones para cada artículo recomendado.

3.6 ALGORITMO DE RECOMENDACIONES

Después de la construcción de los vectores de consumo de clientes y de las reglas de asociación, ambos se agrupan para generar una lista de recomendaciones con una calificación de acuerdo a una función que toma en cuenta una serie de factores que explican a continuación.

La función que asigna a los diferentes tipos de vínculos valores numéricos parte de un diseño propuesto originalmente para la compañía Safeway UK, en la cual Lawrence (2001) se basa en la taxonomía para proponer una heurística que asigna valores de acuerdo al tipo de vínculo, el nivel en la jerarquía de productos y el enfoque por el tipo de relación entre departamentos.

Posteriormente un grupo de especialistas de IBM USA dirigidos por Baragoin (2001) realizan y documentan un ajuste a los valores con el auto-vínculo (E), a nivel clase en una auto relación (Tabla 3.3), buscando asignar una relación más depurada a las relaciones entre productos, evitando así recomendar un producto que ya ha sido previamente adquirido por el cliente.

Para este trabajo se toma la heurística propuesta para Safeway, y la mejora realizada por IBM, además se introduce un ajuste adicional en el vínculo fuerte intra-departamental a nivel clase (B) (Tabla 3.3), que es una mejora dentro de la heurística, y que busca favorecer con un ligero valor adicional a las oportunidades entre departamentos, por encima de las oportunidades en un mismo departamento. En otras palabras, se asigna una calificación más alta a los vínculos entre clases de diferentes departamentos, y una más baja a los vínculos entre clases de un mismo departamento, con el fin de fomentar la diversidad de productos y las ventas cruzadas (Tabla 3.8).

Por otra parte, existe un tipo de vínculo que, a pesar de estar documentado de manera general en el diseño de Safeway UK, no ha sido incluido en la heurística utilizada en este trabajo. Este tipo de vínculo es el relacionado con el diseño manual de reglas de asociación, su potencial podría ser importante para la promoción de ciertas combinaciones de productos.

Para su incorporación se propone un valor heurístico equivalente al de los vínculos fuertes con potencial de cross-selling (ventas cruzadas), lo cual implica otorgarle el valor más alto disponible en la escala empleada. Garantizando se esta manera que tenga un valor elevado y por tanto la recomendación será realizada para todos los clientes, puesto que la heurística posee la capacidad de discernir quienes tienen la propensión al consumo más elevada, dicha capacidad se apoya en los enfoques de filtrado contextual y colaborativo.

La tabla 3.8 resume la heurística y su evolución, para la asignación de calificaciones (valores) a cada producto a ser recomendado de acuerdo al tipo de vínculo, el nivel de productos en la jerarquía, en tipo de relación o enfoque, y finalmente para el caso que se está desarrollando la combinación de todos los factores anteriores favoreciendo las recomendaciones de productos pertenecientes a un departamento diferente al consumido.

Vínculo	Tipo de vínculo	Nivel	Enfoque	Safeway UK	IBM USA	Modificación México
A	Vínculo fuerte	Clase	Ínter-departamental	1.00	1.00	1.00
B	Vínculo fuerte	Clase	Ínter-departamental	1.00	1.00	0.75
C	Vínculo taxonómico	Clase	Ínter-departamental	0.50	0.50	0.50
D	Vínculo débil	Depto.	Ínter-departamental	0.25	0.25	0.25
E	Auto-vínculo	Clase	Relación autógena	0.00	-1.00	-1.00
F	Vínculo diseñado	Clase	Relación exógena	NA	NA	1.00

Heurística de Valores para las Reglas Asociación

Tabla 3.8

Una vez asignados los valores a los productos a ser recomendados, se suma la marca en caso de ser posible, sin olvidar que ésta aplica siempre que se trate de un producto perteneciente a cualquier departamento de vestir, y que exista en el departamento-clase que se está recomendando. En el caso de que la marca favorita no exista en la clase recomendada entonces se asigna una marca afín a la predilecta, en línea con la matriz de marcas en cuanto a moda y precio.

Todos los componentes del proceso de la Generación de Recomendaciones Personalizadas se han descrito previamente, desde la construcción del Data Mart, la generación de un modelo fuente para la Minería de Datos, el proceso de minado de datos con el algoritmo de Reglas de Asociación, la construcción de perfiles, el cruce de información de consumo y relaciones entre productos, la asignación de calificaciones por medio de la heurística y finalmente el empleo de la marca favorita. Lo que continúa es la sistematización del proceso, que se lleva a cabo en la sección siguiente.

3.7 SISTEMATIZACIÓN

La sistematización de los procesos descritos no es una tarea simple, es necesario desarrollar los componentes requeridos y conectar pasos a paso cada uno, con el fin de que cada uno de los clientes obtenga una lista única de recomendaciones, de manera periódica. En el capítulo 2 se construyó el Data Mart y el modelo agregado, en este mismo capítulo ya se realizó el Minado de Datos que produjo las asociaciones entre productos, ahora se automatizan los procesos restantes y se conectan entre ellos.

Los procesos se desarrollan en SQL (Structured Query Language) que es el lenguaje estándar de comunicación con base de datos, a través de SQL dinámico que permite una programación de consultas (Query) no fijas con el fin de recuperar información de la base de datos, así en cada paso se desarrolla un “Query” que genera información de salida útil para los siguientes pasos.

Las recomendaciones se generan mensualmente, pero no todos los procesos deben ejecutarse cada mes, por ejemplo las reglas de asociación se calculan anualmente, ya que difícilmente cambiarían en un tiempo menor, mientras que la matriz de marcas y las marcas favoritas se actualizan semestralmente.

A continuación se muestra el proceso completo con todos los pasos hasta llegar a la generación de las listas de recomendaciones para cada individuo.

Los pasos a seguir se muestran en la figura 3.7, se describen de manera general a continuación, y éstos pueden consultarse a detalle en el anexo B.

1. PASO MM. Matriz de Marcas.

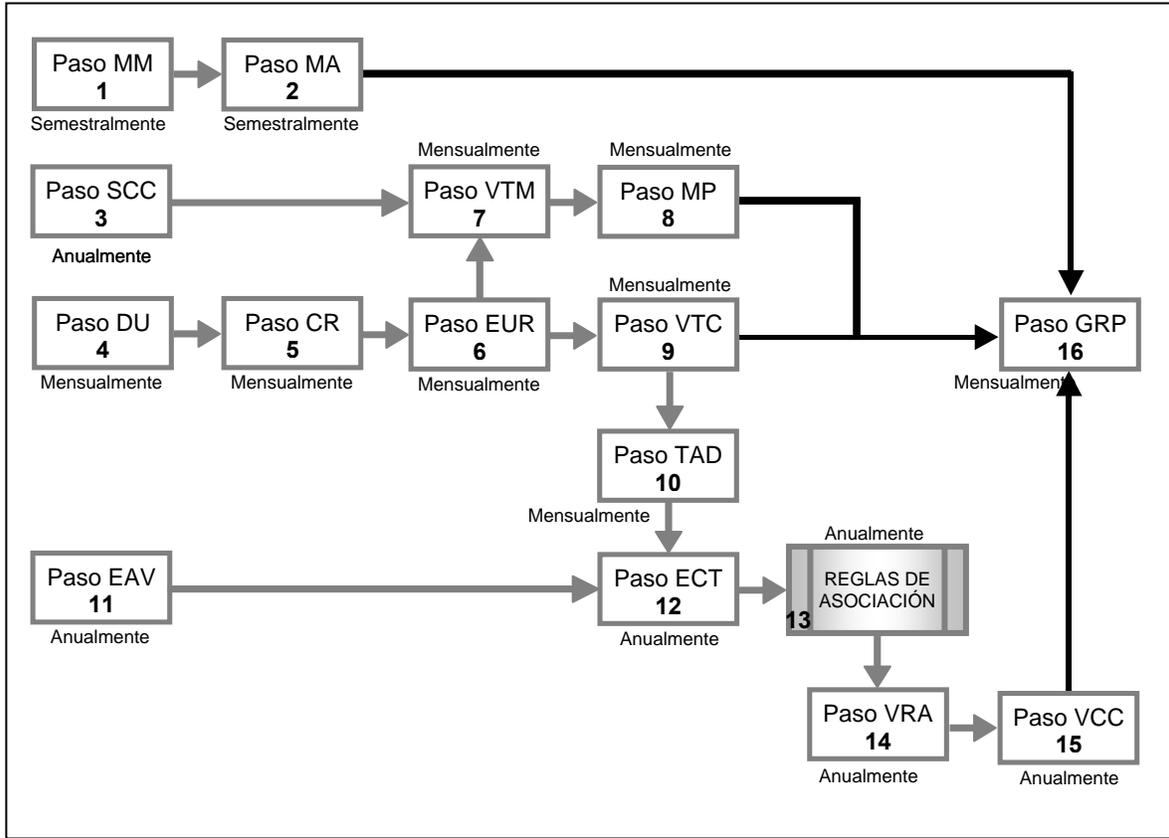
En este paso se almacena la información de la matriz de marcas, de la cual la organización hace una actualización periódica. Este proceso se realiza semestralmente.

2. PASO MA. Marca Afín.

Este paso asigna todas las marcas afines a cada marca para todos los departamentos unificados posibles. Este proceso se realiza semestralmente.

3. PASO SCC. Categorías CRM.

Almacena la información de las categorías generadas para clasificar a los departamentos unificados en una de ellas. Este proceso se realiza anualmente.



Pasos para la Generación de Recomendaciones Personalizadas

Figura 3.7

4. PASO DU. Departamento Unificado.

Almacena la información de los departamentos unificados generados para clasificar los departamentos. Este proceso se realiza mensualmente.

5. PASO CR. Clase Resuelta.

Genera los datos de los productos, los cuales son compuestos por departamento unificado y clase, tanto en claves como en texto. Este proceso se realiza mensualmente.

6. PASO EUR. Estilo Unificado Resuelto.

Unifica los estilos de los productos de tal manera que se consolidan a este nivel de la taxonomía. Este proceso se realiza mensualmente.

7. PASO VT. Vector Tarjeta Marca.

Calcula el gasto relativo por marca para cada tarjeta. Este proceso se realiza mensualmente.

8. PASO MP. Marca Preferida.

Obtiene la marca favorita del cliente, que es aquella en la que más ha gastado el cliente. Este proceso se realiza mensualmente.

9. PASO VTC. Vector Tarjeta Clase.

Calcula el gasto relativo por departamento unificado y clase para cada tarjeta. Este proceso se realiza mensualmente.

10. PASO TAD. Tarjetas de Alta Diversidad.

Selecciona aquellas tarjetas que tienen alta diversidad de compras, se obtiene cuando el cliente ha comprado en 5 o más categorías diferentes, con el gasto de 50% o menos en cada una de ellas. Este proceso se realiza mensualmente.

11. PASO EAV. Estilos de Alto Valor.

Filtra aquellos productos cuyo valor sea mayor o igual a \$100, utilizando el nivel estilo de la taxonomía. Este proceso se realiza anualmente.

12. PASO ECT. Evento Compra Trimestral

Calcula los departamentos unificados donde el cliente gastó durante el año para cada trimestre, para las tarjetas de alta diversidad y los estilos de alto valor.

Calcula las clases-departamento unificado donde el cliente gastó durante el año para cada trimestre, para las tarjetas de alta diversidad y los estilos de alto valor.

Este proceso se realiza anualmente.

13. REGLAS DE ASOCIACIÓN.

Almacena las relaciones de productos generadas con el algoritmo de Minería de Datos en la base de datos, con el soporte, la confianza y el lift. Este proceso se realiza anualmente.

14. PASO VRA. Vistas de Reglas de Asociación.

Clasifica las reglas por tipo de vínculo en intra-departamentales e Inter-departamentales, que son los escenarios usados para la generación de éstas. Este proceso se realiza anualmente.

15. PASO VCC. Vector Clase Clase.

Asigna la calificación de para cada una de las reglas de asociación de acuerdo a la heurística propuesta. Este proceso se realiza anualmente.

16. PASO GRP. Generación de Recomendaciones Personalizadas.

Genera las recomendaciones a nivel clase para cada cliente, de acuerdo a su vector de consumo, a las relaciones entre productos y la heurística de calificaciones, también asigna una marca favorita o afín en la recomendación si es posible. Este proceso se realiza mensualmente.

Una vez que el proceso se ha sistematizado e implementado, se tienen lista de recomendaciones para todos los clientes, generadas de forma mensual, a continuación se revisan algunas de los resultados obtenidos.

3.8 RESPUESTA DEL MODELO

En términos generales, el proceso es capaz de producir una lista de recomendaciones para cada tarjeta de crédito que haya comprado algo en el último año, a partir del perfil de hábitos de consumo se relacionan productos derivados del Algoritmo de Asociaciones, se asigna una calificación dependiendo del tipo de relación, como punto final el producto recomendado se complementa si es permisible con la marca. A continuación se examinan algunas de las listas de recomendaciones, así como los componentes de las mismas.

3.8.1 Elementos de la recomendación

A continuación se presentan los elementos que forman parte de las listas de recomendaciones de consumo personalizadas, estos son: tarjeta, calificación, producto y marca.

- **Tarjeta.**

Representa el número de tarjeta de crédito hacia la cuál se desea proponer el producto recomendado, identifica al cliente, es a partir de ella que se establece el vector que representa el perfil de consumo, y es a partir de ella también que se generan los tickets que son fuente para la procesamiento del algoritmo de Minería de Datos.

- **Calificación (Score).**

El espacio de afinidades de productos se encuentra densamente poblado, prácticamente cualquier preferencia puede producir recomendaciones y generar listas exhaustivas de los productos disponibles. La importancia de las recomendaciones se encuentra ordenada de acuerdo a la calificación de propensión al consumo (Score), generada con la heurística, debido a esta situación no es recomendable profundizar en el número de elementos dentro de cada lista puesto que la calificación puede decaer de manera importante, es por esta razón que el número de productos de la lista de cada lista se ha limitado a 10.

- **Productos**

Los productos recomendados son el resultado del procedimiento de filtrado contextual y colaborativo, después de la búsqueda de relaciones derivadas del minado de datos. Por la manera en que fueron seleccionadas las reglas de negocio en el proceso, el nivel de la recomendación, es decir el producto, se representa por una clase asociada a un departamento unificado.

- **Marca**

La matriz de marcas comprende mercancías que tienen que ver con artículos de vestir, involucrando tendencias de moda, ya sean estilos clásicos o contemporáneos. Por esta razón no todos los productos recomendados pueden ser asociados por una sugerencia de marca.

Para determinar que marca a recomendar, se tienen en cuenta varios elementos como la preferencia de marca que ha mostrado el cliente en los últimos doce meses de compra. Tomando como fundamento la matriz de marcas puede establecerse un cierto nivel de precio y moda, apoyándose después en su estructura se pueden derivar marcas afines, con el fin de encontrar otras marcas que pueden proponerse cuando la marca preferida no está disponible en la clase recomendada.

Los siguientes planteamientos van en el sentido de establecer un conjunto de alternativas que permitan incorporar los resultados del proyecto en diferentes escenarios buscando favorecer diferentes estrategias en la tienda.

En este punto concluye el análisis, diseño y la implementación del sistema de recomendaciones, lo que continua es revisar los resultados y llevar a cabo acciones con los mismos, en la siguiente sección se revisan algunas lista de recomendaciones obtenidas del proceso.

3.8.2 Ejemplos de listas de recomendaciones

A continuación se presentan seis ejemplos de listas de recomendaciones para clientes elegidos aleatoriamente, estas listas contienen 10 productos con las mejores calificaciones, es decir para los cuales la propensión al consumo es más altas, en algunos casos los productos están acompañados de una marca que puede ser la favorita o una afín.

Se puede observar que las listas tienen diferentes estilos que reflejan de alguna manera la personalidad de los clientes, desde mujeres de poder adquisitivo medio y alto (derivado de las marcas), que son los casos de las tarjetas 1 (Tabla 3.9), 3 (Tabla 3.11) y 6 (Tabla 3.14); enfocadas el esparcimiento que es el caso de la tarjeta 2 (Tabla 3.10); y las que tienen una orientación a la familia y el hogar, que son los casos de las tarjetas 4 (Tabla 3.12) y 5 (Tabla 3.13).

Tarjeta	Score	Producto Recomendado	Marca Recomendada
1	0.010209	DAMA // Falda	CHARME
	0.008674	DAMA // Traje Sastre	CHARME
	0.008674	DAMA // Conjunto	CHARME
	0.008674	DAMA // Gabardina	CHARME
	0.007954	PETITS // Saco	CHARME
	0.007823	AREA JOVEN DAMA // Falda	
	0.007413	JOYERIA Y RELOJES // Fina	
	0.007132	ACCESORIOS DAMA // Chal y Bufanda	ARIES
	0.006996	JOYERIA Y RELOJES // Fantasía	
	0.006984	SPRINGFIELD // Camisas	

Lista de recomendaciones Tarjeta 1

Tabla 3.9

Tarjeta	Score	Producto Recomendado	Marca Recomendada
2	0.022468	COMPUTO // Impresora	
	0.009366	LOUIS VUITTON // Calzado Dama	LOUIS VUITTON
	0.009366	LOUIS VUITTON // Calzado Caballero	LOUIS VUITTON
	0.009366	LOUIS VUITTON // Chal y Bufanda	LOUIS VUITTON
	0.007793	REGALOS EXCLUSIVOS // Relojes	
	0.007793	REGALOS EXCLUSIVOS // Escultura- Portaret	
	0.007793	REGALOS EXCLUSIVOS // Velas	
	0.007737	SALVATORE FERRAGAMO ZAP // Zapato Sport Dama	
	0.004084	DEPORTES // Campismo y Pesca	
	0.004084	DEPORTES // Natacion y Buceo	

Lista de recomendaciones Tarjeta 2

Tabla 3.10

Tarjeta	Score	Producto Recomendado	Marca Recomendada
3	0.009215	LOUIS VUITTON // Calzado Dama	
	0.009215	LOUIS VUITTON // Calzado Caballero	
	0.009215	LOUIS VUITTON // Chal y Bufanda	
	0.007367	SALVATORE FERRAGAMO ZAP // Zapato Sport Dama	
	0.006823	DAMA // Sweater	CHARME
	0.006192	DAMA // Traje Sastre	CHARME
	0.006192	DAMA // Conjunto	CHARME
	0.006192	DAMA // Chamarra	CHARME
	0.005904	CALZADO DAMA // Confort	LAURENT
	0.005386	CALZADO DAMA // Gala	LAURENT

Lista de recomendaciones Tarjeta 3

Tabla 3.11

Tarjeta	Score	Producto Recomendado	Marca Recomendada
4	0.007857	SALON INTERNACIONAL // Gabardina	ELIO BERHANYER
	0.007857	SALON INTERNACIONAL // Traje Sastre	ELIO BERHANYER
	0.007857	SALON INTERNACIONAL // Pantalón	ELIO BERHANYER
	0.007836	BLANCOS Y BAÑOS // Closet	
	0.007836	BLANCOS Y BAÑOS // Mantelería	
	0.005316	SALVATORE FERRAGAMO ZAP // Zapato Sport Dama	
	0.005105	BOUTIQUE DAMA // Vestido	ANNE KLEIN
	0.005105	BOUTIQUE DAMA // Playera	ANNE KLEIN
	0.005105	BOUTIQUE DAMA // Traje de baño	ANNE KLEIN
	0.005105	BOUTIQUE DAMA // Capa	ANNE KLEIN

Lista de recomendaciones Tarjeta 4

Tabla 3.12

Tarjeta	Score	Producto Recomendado	Marca Recomendada
5	0.145135	COCINA INTEGRAL Y BAÑO // Sobre medida	
	0.017449	PISOS Y TAPETES // Tapete hecho maquina	
	0.017449	PISOS Y TAPETES // Pisos	
	0.017449	PISOS Y TAPETES // Tapete hecho a mano	
	0.00648	BASICOS DESCANSO // Reclinables	
	0.00648	BASICOS DESCANSO // Literas y Camas	
	0.00648	BASICOS DESCANSO // Sofas cama	
	0.00355	MUEBLES // Tapizados tela	
	0.00355	MUEBLES // Recamaras	
	0.00355	MUEBLES // Muebles oficina	

Lista de recomendaciones Tarjeta 5

Tabla 3.13

Tarjeta	Score	Producto Recomendado	Marca Recomendada
6	0.039104	PIELES // Chaqueta	
	0.039104	PIELES // Falda	
	0.039104	PIELES // Chaleco	
	0.039104	PIELES // Capa	
	0.039104	PIELES // Chamarra	
	0.013794	SALON INTERNACIONAL // Vestido	ST. JOHN
	0.013794	SALON INTERNACIONAL // Chamarra	ST. JOHN
	0.013794	SALON INTERNACIONAL // Bolsas	ST. JOHN
	0.013794	SALON INTERNACIONAL // Casual	ST. JOHN
	0.013794	SALON INTERNACIONAL // Vestir	ST. JOHN

Lista de recomendaciones Tarjeta 6

Tabla 3.14

Una vez que se han generado las listas de recomendaciones, el siguiente paso es darlas a conocer al cliente y medir el efecto. Esto depende entre otras cosas del medio de comunicación utilizado y del tiempo en que se realiza la recomendación.

Tal vez el cliente se interese en el producto, le llame la atención y desee comprarlo, pero eso no significa que lo hará de inmediato, lo cual se considera en el siguiente capítulo.

CONCLUSIONES

En este capítulo se desarrolla el proceso de generación de listas personalizadas de recomendaciones para cada cliente.

En primer lugar se examinan los algoritmos de Minería de Datos, sus características y posibles usos, se selecciona el método más adecuado para el desarrollo del proceso apegado al software disponible, se lleva a cabo el proceso de minado de datos hasta encontrar el resultado mostrado, posteriormente se revisan las relaciones surgidas entre productos con sus valores correspondientes, el siguiente paso es crear los perfiles de consumo para cada cliente, para luego integrar ambos perfiles y relaciones con base en una heurística que califica cada producto a recomendar para distinguir entre los productos más o censo recomendables, y al final de ser posible se adiciona la marca favorita o afín, el producto final de proceso son lista de recomendaciones de cada cliente.

Cabe señalar que los procesos desarrollados están fundamentados en una tecnología robusta de base de datos, así como un software de Minería de Datos.

La Minería de Datos prueba ser una herramienta poderosa para la generación de información que de otra manera sería muy difícil de conocer, pero también es evidente que si se combina con otras técnicas y procesos, el resultado puede ser asombroso. Sin las relaciones entre productos encontradas a través del Algoritmo de asociaciones las recomendaciones difícilmente podrían tener ese sentido de veracidad, y consecuentemente provocar un efecto notorio en el consumo de los clientes. En el siguiente capítulo se revisan las aplicaciones de esta información, las cuales pueden ser en diferentes sentidos y a través de diferentes medios dirigidos, también se hacen mediciones para conocer el efecto de las mismas.

Capítulo IV

Aplicación

INTRODUCCIÓN

El ofrecer al cliente el producto adecuado en el tiempo idóneo, con el precio apropiado y con un buen servicio, puede resultar un instrumento muy poderoso para la retención y desarrollo del cliente. El cliente ha evolucionado a través del tiempo, ya no existen más los consumidores cautivos que se conformaban con el cumplimiento de las necesidades, ahora el exceder las expectativas es más y más normal, por lo cual el reto de cualquier compañía es cada vez más grande debido a la intensa competencia y a la globalización.

Este trabajo centra sus esfuerzos en identificar las necesidades del consumidor, muchas veces sin que éste las conozca de manera explícita. Es importante que el cliente reciba sugerencias de productos que le resulten atractivos y no de aquellos que están fuera de sus necesidades y expectativas. Cuando un cliente recibe una recomendación de un producto que le resulta atractivo o interesante, entonces el cliente se sentirá entendido y conocido, muy probablemente se rebasen sus expectativas y se alcance un paso más allá.

El siguiente paso después de generar las listas de recomendaciones para cada cliente, es darles a conocer los productos contenidos en las listas, es tan importante la forma como el medio de hacerlo, finalmente conocer el efecto causado en el consumo antes y después del uso de las recomendaciones personalizadas, que se revisan en el presente capítulo.

El capítulo está dividido de la siguiente manera, en la primera parte se examinan los posibles caminos en el uso de las listas de recomendaciones, la segunda parte trata sobre los medios de contacto con el cliente, en la tercera sección se propone el uso alternativo de algunos datos derivados del proceso, y en el último punto se llevan a cabo pruebas en el uso de las recomendaciones, junto con una medición de la efectividad en su uso.

4.1 APLICACIONES

El problema se desarrolla en una cadena de tiendas departamentales, por lo que existen productos diversos. Los clientes a los que se les harán recomendaciones son todos aquellos que cuentan con una tarjeta de crédito de la tienda y con compras en el último año, es decir 332,470 tarjetahabientes, en cuanto a los productos, se toman 241 departamentos unificados y 1,815 clases. Esto significa que se pueden recomendar más de mil ochocientos productos a más de trescientos mil clientes.

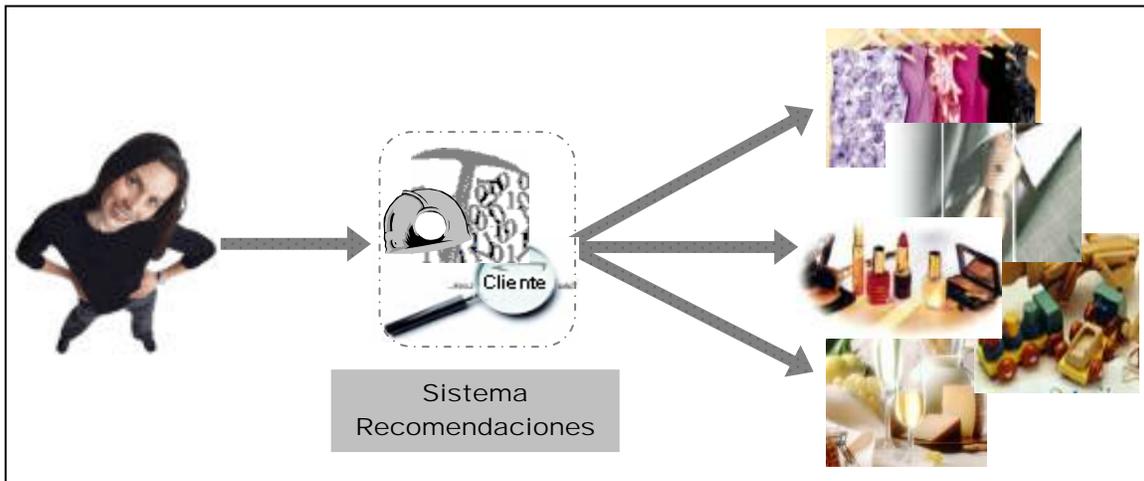
Las recomendaciones fueron creadas con base en Reglas de Asociación de Minería de Datos para conectar productos y consumo de clientes por medio de sus hábitos de compra.

Se desea establecer un conjunto de alternativas para incorporar los resultados en diferentes escenarios, buscando favorecer diferentes estrategias de la compañía. Los diferentes esquemas o alternativas de aplicación no son excluyentes y puede combinarse con los medios de contacto.

A continuación se muestran los enfoques en los que se pueden aplicar las listas de recomendaciones.

4.1.1 Enfoque natural

La aplicación natural del sistema es en el sentido cliente → producto (Figura 4.1), es decir, dado que se identifica a un cliente se le sugiere una lista de productos, los cuales deben resultarle atractivos o interesantes.



Relación Cliente – Producto

Figura 4.1

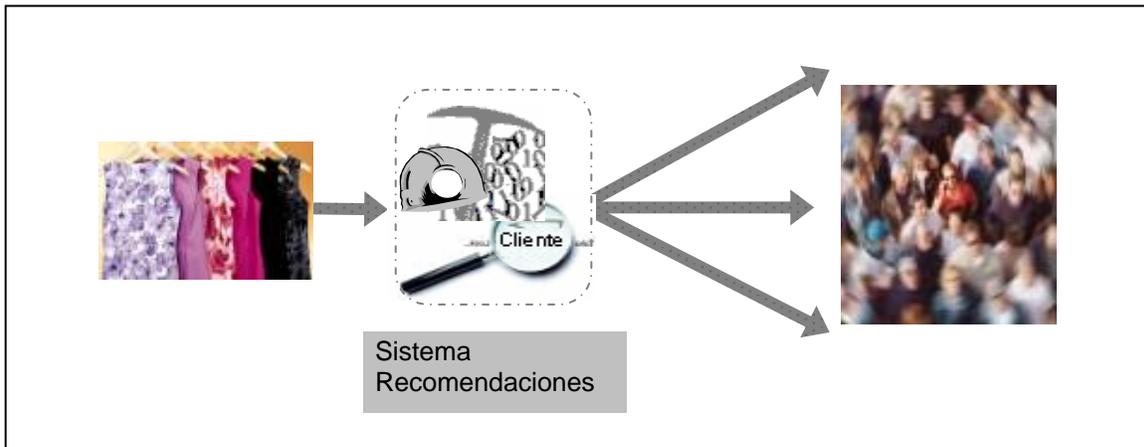
Lo importante de este enfoque es la manera en como se dan a conocer las recomendaciones al cliente, el momento, la cantidad de productos que se dan a conocer al mismo tiempo y la frecuencia.

Las listas se limitaron a un máximo de 10 artículos, se generan de manera mensual y lo más relevante es que cada producto tiene una calificación. Debe tenerse cuidado de no saturar al cliente invitándolo a adquirir diez artículos mensualmente, ya que puede llegar a ser muy molesto. Una estrategia de máximo tres artículos por cliente mensualmente es de vital importancia, un elemento fundamental es la calificación, ya que con base en ésta se pueden elegir los productos más recomendables para cada caso.

Este enfoque implica una comunicación directa y personal con el cliente, que se propone sea directamente en el punto de venta, cuando el cliente realice una transacción y el vendedor desliza la tarjeta, o bien cuando el cliente consulta la página de Internet y se identifica con su número de tarjeta.

4.1.2 Enfoque inverso

Otro camino a seguir es invertir el enfoque natural utilizando el sentido producto → cliente (Figura 4.2), en este caso se asume que se tiene establecido un producto que se desea promocionar o bien se rastrea el producto más recomendado, en general o por departamento, posteriormente se seleccionan aquellos clientes que presentan las calificaciones de propensión al consumo más altas. De esta manera se recomienda al cliente una cosa a la vez, es decir, se encuentra un grupo de clientes que sean altamente propensos a comprar un producto y de manera masiva se les hace llegar la recomendación.



Relación Producto – Cliente

Figura 4.2

Este enfoque involucra una comunicación masiva con el cliente pero con un mensaje personalizado, para comunicar de esta manera se pueden elegir diferentes medios de contacto masivos pero individualizados, los cuales se revisan en la siguiente sección.

4.2 MEDIOS DE CONTACTO

Para hacer llegar las sugerencias al cliente existen varios medios de contacto, que pueden llegar a ser exitosos en diferentes niveles y que implican diferentes costos, desde personas en los puntos de venta (vendedores), correo normal en diferentes modalidades, telemarketing y correo electrónico, a continuación se presenta de manera puntual los medios de contacto mencionados.

4.2.1 Punto de venta

Este es el medio de comunicación más efectivo, interactuar con el cliente frente a frente cuando éste se encuentra en la tienda. El momento en que el vendedor desliza la tarjeta del cliente representa el más importante, ya que es el momento de identificarlo y de obtener la mejor recomendación de ese momento. Si además de sugerirle al cliente un producto importante o interesante para él se le da un buen servicio, se reúnen varios elementos que generan condiciones propicias para obtener una respuesta favorable.

La utilización de las listas de recomendaciones no solo persigue un impacto económico, también el de generar un sentido de cercanía o conocimiento del cliente, lo cual finalmente puede incrementar su lealtad, que es cada vez más difícil de conseguir en estos tiempos.

La funcionalidad de contar automáticamente con las recomendaciones de todos clientes en cada punto de venta, implica desarrollar una solución informática simple para su funcionamiento en todas las terminales de venta. A continuación se presenta una alternativa para el diseño de ésta, que considera dos tipos de procesos:

- **Proceso mensual.**

Es el encargado de extraer las recomendaciones de cada mes para cada cliente, es decir la transferencia de los 3 artículos a una base de datos que pueda ser leída por los puntos de venta, con el número de tarjeta, el artículo recomendado, su calificación, una fecha control de frecuencia y un indicador de sí ya se ha dado a conocer la recomendación o no.

- **Proceso en línea.**

Implica modificar el operativo, es decir el software en el punto de venta, para incluir una consulta a la estructura que contiene los tres productos, para que al momento de deslizar la tarjeta para pagar, se envíe el número, se haga una consulta de las recomendaciones y se despliegue la recomendación en la pantalla, de manera que el vendedor lea la recomendación y se la comunique de manera verbal al cliente. Este proceso debe tardar sólo segundos para que sea eficiente.

4.2.2 Telemercadeo

Este medio de comunicación es cada vez más usado, dentro de la organización usualmente se emplea para dar a conocer promociones o ventas especiales a un grupo de clientes seleccionados previamente.

Si el cliente recibe una llamada donde lo invitan a comprar un producto de su interés, el cual adicionalmente está en promoción con algún descuento y/o con mensualidades sin intereses y/o se obsequia un regalo, el impacto que puede generar es muy bueno.

Para este tipo de marketing, la línea a seguir son los artículos pertenecientes a las promociones, con base en esta información se crea una lista de clientes con recomendaciones de estos productos, con la opción de elegir productos por calificación, posteriormente se llevan a cabo las llamadas.

4.2.3 Correo tradicional

Este medio de comunicación es el más antiguo de todos pero no por eso es ineficiente, y puede tomarse en dos sentidos: estados de cuenta y propaganda impresa.

- **Estado de cuenta.**

El cliente que recibe su estado de cuenta impreso de manera mensual, es sujeto a recibir más información conjuntamente, que puede ser propaganda impresa dentro del sobre o bien como un mensaje personalizado impreso en el estado de cuenta.

- **Envíos de propaganda.**

Otro tipo de comunicación es enviar separadamente del estado de cuenta folletos, postales, invitaciones, catálogos, etc. Igualmente se selecciona a un grupo de clientes con intereses compatibles referentes a la promoción

4.2.4 Correo electrónico

No todos clientes cuentan con una dirección de correo electrónico, pero cada día se incrementa el número de personas con e-mail, además de que aunque el cliente tenga uno no necesariamente está registrado en la base de datos.

Es un medio de comunicación poderoso, ya que permite diseñar mensajes completamente individualizados que den a conocer las recomendaciones personalizadas, que podrían ser acompañados de una funcionalidad para agregar una liga que permita conocer más información al

respecto, de esta manera se puede dar un seguimiento puntual a las acciones del cliente desde la recepción del correo.

4.2.5 Página en Internet

La tienda cuenta con una página en Internet, a través de la cual los clientes pueden realizar compras como en una tienda normal. Este es otro medio de contacto importante y que crece día con día.

La manera de dar a conocer las recomendaciones es similar a un punto de venta, cuando un cliente teclea su número de tarjeta, en ese momento un proceso envía el número que realiza la consulta, para que en un tiempo corto se muestre una ventana con las recomendación.

4.3 APLICACIONES COMPLEMENTARIAS

Algunas aplicaciones complementarias surgen a partir de la información creada durante el proceso de generación de recomendaciones.

- **Marcas preferidas.**

Este dato por si solo es interesante, una aplicación extra es usarlo directamente en la promoción de una marca, y conjuntamente con otro tipo de información puede mejorar la planeación de diferentes estrategias de la compañía, apoyándose en la matriz de marcas.

- **Vector de tarjeta a nivel clase.**

Esta información contiene el detalle de lo que el cliente ha comprado en el último año, con un nivel de agregación tan específico que podría vincularse con preferencias, diversidad, lealtad a las marcas y consumo.

- **Marcas afines.**

Esta información permite impulsar marcas ya que puede usarse en el mismo sentido que la marca preferida, también se puede analizar la oferta, demanda y rentabilidad en cuanto a moda y precio apoyándose en la matriz de marcas.

4.4 PRUEBAS

La implementación del proceso no es simple necesita tiempo y recursos, debido a esto el desarrollo en el punto de venta está aún en proceso pero se tienen los otros tipos de comunicación, hasta el momento se ha aplicado el telemarketing con enfoque inverso.

La manera de operar de la tienda es planear promociones y darlas a conocer a los clientes, normalmente se selecciona a un grupo de personas que compraron en la misma promoción el año anterior, o bien que han comprado durante los últimos meses productos pertenecientes a la promoción.

La aplicación que se muestra a continuación se llevó a cabo con una campaña relacionada con muebles, la cual se ha promovido por varios años, además contemplar a los clientes que normalmente se tomarían, se agregó un grupo de clientes cuyas recomendaciones (las 3 primeras) son de productos relativos a la promoción.

Normalmente se hacen campañas por zonas y se llama a clientes con dirección dentro de la zona, para este caso se tienen 3 zonas, una de las cuales contempla una tienda relativamente nueva con pocos meses en funcionamiento (Zona 3), y es la primera vez que se realiza esta promoción, por lo cual no existe un grupo de clientes definido para el telemarketing y se decide llamar a aquellos clientes cuyo límite de crédito es alto, bajo el supuesto que tienen la capacidad de compra de productos relacionados con muebles.

Después de realizar el telemarketing y evaluar el efecto en causado en los clientes, es decir el porcentaje de clientes que compraron en la promoción, puede observarse que tuvo un efecto positivo para las zonas 1 y 2 y excelente para la zona 3 (Tabla 4.1)

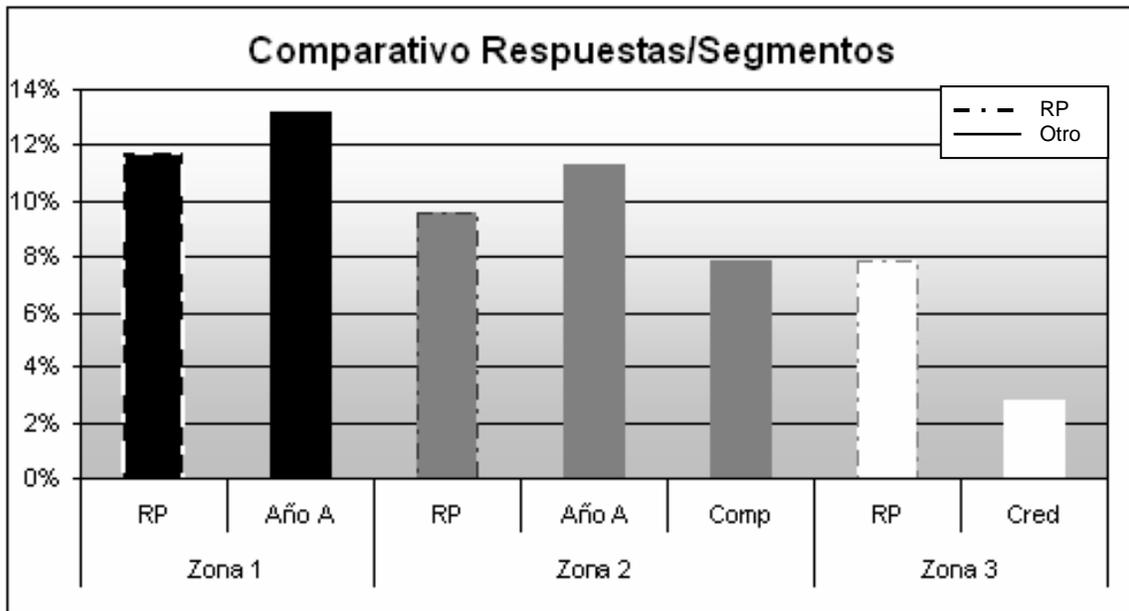
Para las zonas 1 y 2 la promoción ya se había realizado durante varios años, se llamó a clientes hicieron compras en el año anterior, es decir clientes que encontraron atractiva la promoción y que resulta más simple ser atraídos nuevamente, con una respuesta de 13.2% y 11.3% respectivamente, para la zona 2 también se contactó a clientes con compras de productos relacionados obteniendo una respuesta de 7.9%, mientras que la respuesta de los clientes con recomendaciones personalizadas fue de 11.7% y 9.6% respectivamente, que es una respuesta muy cercana a clientes ya convencidos por una experiencia previa, lo cual puede decirse que es un efecto positivo y difícil de alcanzar con criterios de tipo más general, incluso con clientes que ya han comprado este tipo de productos, que es el caso de la zona 2.

La zona 3 representa un reto complicado, ya que no se cuenta con historia previa, sólo con información de algunos meses, en este caso es evidente que la Minería de Datos y los hábitos de compra pueden causar un efecto importante, y que criterios generales como es la capacidad de compra no necesariamente significa atracción de compra, la respuesta para límites de crédito alto es de 2.8% mientras que la de recomendaciones es de 7.9%, lo cual es una respuesta más que favorable, sin ésta sólo se hubiera obtenido el 2.8%, lo cual puede observarse de manera gráfica en la figura 4.3.

Zona 1		Zona 2		Zona 3	
Segmento	Respuesta	Segmento	Respuesta	Segmento	Respuesta
Recomendaciones Clientes con recomendaciones en productos de la promoción, en los 3 primeros lugares.	11.7%	Recomendaciones Clientes con recomendaciones en productos de la promoción, en los 3 primeros lugares.	9.6%	Recomendaciones Clientes con recomendaciones en productos de la promoción, en los 3 primeros lugares.	7.9%
Año Anterior Clientes con compras promoción año anterior.	13.2%	Año Anterior Clientes con compras promoción año anterior.	11.3%	Crédito alto Clientes con Límite de crédito alto.	2.8%
		Compras Clientes con compras de productos de la promoción en el último año.	7.9%		

Efecto en compras de clientes

Tabla 4.1



Comparativo Respuestas por Zonas

Figura 4.3

Se han llevado a cabo más telemercadeos donde se han incluido clientes por Recomendaciones Personalizadas, al comparar con selecciones de clientes que no han tenido ingerencias previas con las promociones, de manera general se ha visto un incremento en la respuesta de entre un 20% y 30% del porcentaje de respuesta normal.

En el futuro se planea probar con todos los medios de contacto y con los dos enfoques de recomendaciones, se está preparando el proceso para que en breve se hagan envíos vía mail tradicional, así como correo electrónico y en el sitio de Internet.

CONCLUSIONES

La información que se genera con el proyecto es muy valiosa, la obtención de las listas de recomendaciones implica un trabajo arduo, el cual se concibió originalmente para sugerir al cliente de manera directa la compra de un producto, lo cual implica que si el cliente ya está comprando es más simple que continúe en la misma línea. También existen otras maneras para lograrlo, se puede seleccionar a un grupo de clientes para los que es recomendable un producto que se desea promocionar. Este trabajo expone una aplicación en este sentido con telemercadeo, y se ha comprobado la efectividad de las recomendaciones. Por otra parte, se planea aplicar en el futuro los otros medios de comunicación.

Después de realizar pruebas con casos reales, es evidente que la utilización de las recomendaciones es de gran ayuda y hace que los esfuerzos se enfoquen a aquellos clientes que pueden generar una mejor respuesta, lo cual se traduce en que los recursos que se invierten traen una mejor respuesta, enfocándose a aquellos clientes con mayor propensión.

Se puede afirmar que la evaluación hecha con el caso real mostrado en este capítulo, es una prueba de que la Minería de Datos conjuntamente con el resto del proceso marcan una diferencia favorable, que se ve reflejada directamente en una respuesta favorable por parte de los clientes y que indirectamente implica mayor gasto de estos así como el incremento de su lealtad.

Conclusiones

La mercadotecnia ha evolucionado a través de los años, mientras en el pasado el objetivo de las compañías era cumplir, actualmente es el sobrepasar las expectativas de los clientes, y no sólo vender un producto, sino obtener fidelidad de los clientes.

La lealtad del cliente, en nuestros días es muy difícil de obtener, sobretodo, en las tiendas de tipo departamental, donde se presenta un reto importante debido a la competencia que existe en cuanto a precios y productos, así como la publicidad para el cliente. Por ello, las empresas han desarrollado distintas estrategias de acercamiento, sin embargo, cada cliente tiene diferentes hábitos de consumo y por tanto diferentes intereses, por esto mismo lo que se le ofrece no puede ser lo mismo para todos, debe existir diferenciación y un acercamiento más personal.

En este sentido, la estrategia de acercamiento al cliente, propuesta en esta tesis, se basó en recomendarle al cliente productos apegados a su estilo de compra o a sus intereses, es decir proporcionarle el producto más idóneo en el momento más adecuado.

Para ello, se generaron recomendaciones personalizadas para cada cliente, de acuerdo con sus hábitos únicos de compra, las afinidades de la cartera de productos ofrecida, y la matriz de marcas de moda y precio. La esencia de las recomendaciones se basa en la Minería de Datos, específicamente en el algoritmo de asociaciones (Market Basket Analysis), donde se analizaron y determinaron los vínculos entre artículos, que tan fuertes o débiles son y la conglomeración de los mismos. Esta información junto con los hábitos de compra de cada cliente, la heurística de tipo de relaciones entre productos, y finalmente la matriz de marcas, da como resultado las lista personalizadas de artículos para cada uno de los clientes.

Para lograr esto se analizó a fondo el entorno de información, del negocio mismo y del consumo de los clientes, se realizó la estructura y explotación de datos, se revisó la ingeniería de información, desde tomar la decisión de crear una base de datos acorde con el objetivo del proyecto, el diseño y construcción del modelo de datos congruente con la metodología que usada para la generación de recomendaciones personalizadas y en los recursos tanto de software como hardware disponibles.

Asimismo, se revisaron los diferentes algoritmos de minería de datos, seleccionando el más adecuado para el desarrollo óptimo del proceso, así como las relaciones surgidas entre productos. Posteriormente se crearon los perfiles de consumo o hábitos de compra de cada cliente, logrando así ofrecer productos a los clientes que tienen más afinidad hacia ellos, es decir tener un contacto de mayor precisión con los clientes.

Se puede afirmar que la Minería de Datos ha probado ser una herramienta importante para conocer las relaciones entre productos, ocultas entre los datos e ignoradas previamente, que con otro tipo de herramientas deben ser conocidas de antemano. Lo cual permite el acoplamiento de

esta información con los hábitos de compra de los clientes, a través de una heurística de tipo de relaciones y finalmente enriquecida con el uso de la matriz de marcas.

Pruebas llevadas a cabo para un grupo de clientes, utilizando la metodología tradicional y las listas de recomendaciones, han mostrado la efectividad de las recomendaciones personalizadas, en algunos casos rebasado las expectativas de compra de los métodos tradicionales. Lo cual refuerza la afirmación de la efectividad de la Minería de Datos en este tipo de problemas.

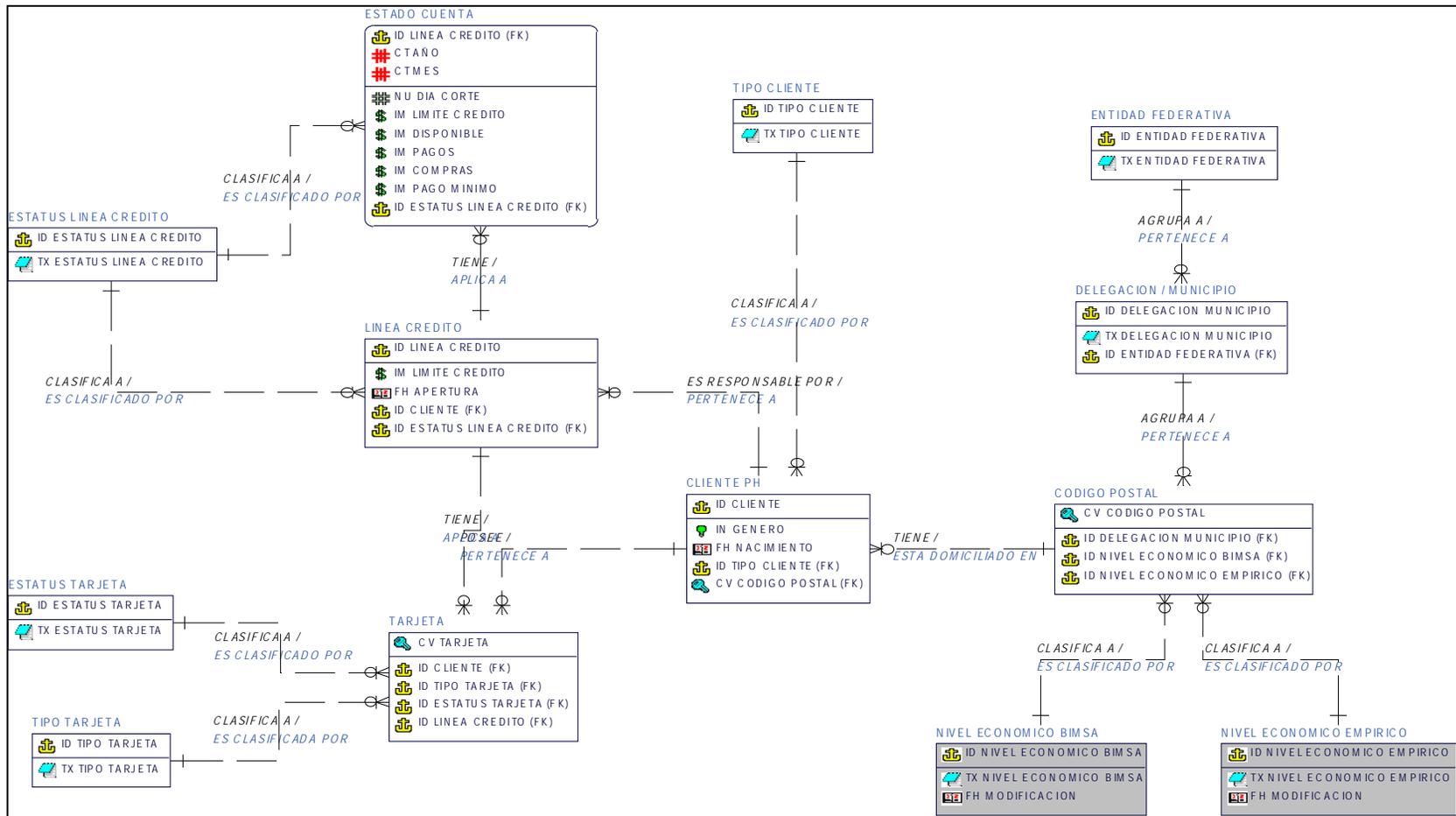
De acuerdo a los resultados obtenidos, se puede señalar que la utilización de la Minería de Datos en el conocimiento de las relaciones entre productos para generar recomendaciones personalizadas, y el ofrecimiento de artículos específicos a un cliente, que sean de preferencia, con base a los patrones encontrados incrementa la probabilidad de que el cliente realice una compra, comparado con el ofrecimiento de productos sin conocer la relación puntual entre ellos, es decir, el ofrecer un producto de interés resulta ser un instrumento poderoso para incrementar las compras de un cliente y por tanto su fidelidad.

Anexo A

Por simplicidad el modelo ER del Data Mart CRM se fragmentó en áreas temáticas, se utiliza un código de colores para la rápida identificación de las fuentes de información.

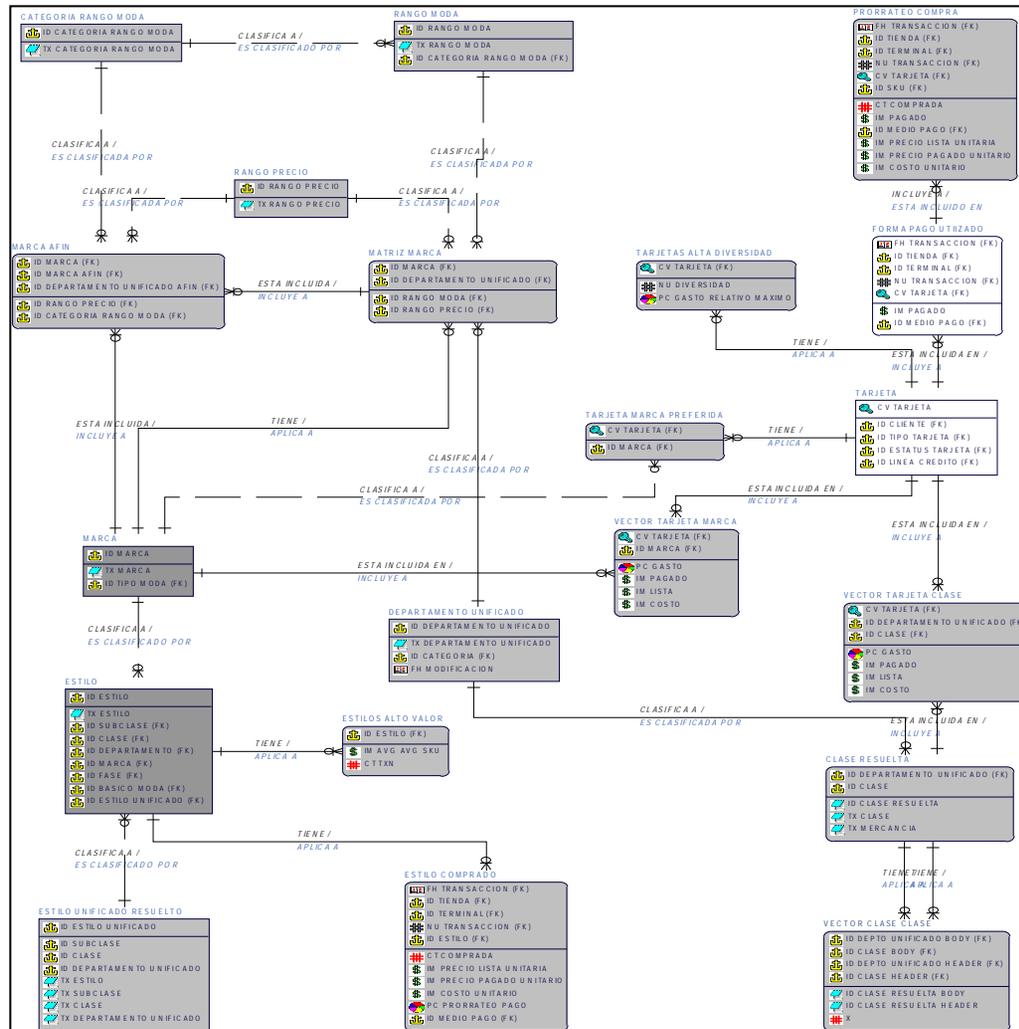
- El color rosa representa la fuente origen Crédito.
Comprende la información de clientes, es decir datos crediticios y demográficos.
- El color verde representa la fuente origen Mercaderías.
Representa la estructura en cuanto a productos, considerando los atributos de los artículos.
- El color amarillo representa información de Promociones.
Contiene la información de promociones, que incluyen los departamentos o productos participantes, así como las fechas de vigencia.
- El color azul representa información propia de CRM.
Incluye la información utilizada directamente en la generación de recomendaciones personalizadas.
- El color gris representa información de estructuras de minado.
Engloba la información que se usa para alimentar el algoritmo de Asociaciones en la herramienta de Minería de Datos.

A continuación se muestran las áreas temáticas del modelo ER.



Modelo ER Data Mart CRM, área temática Clientes
Figura A1

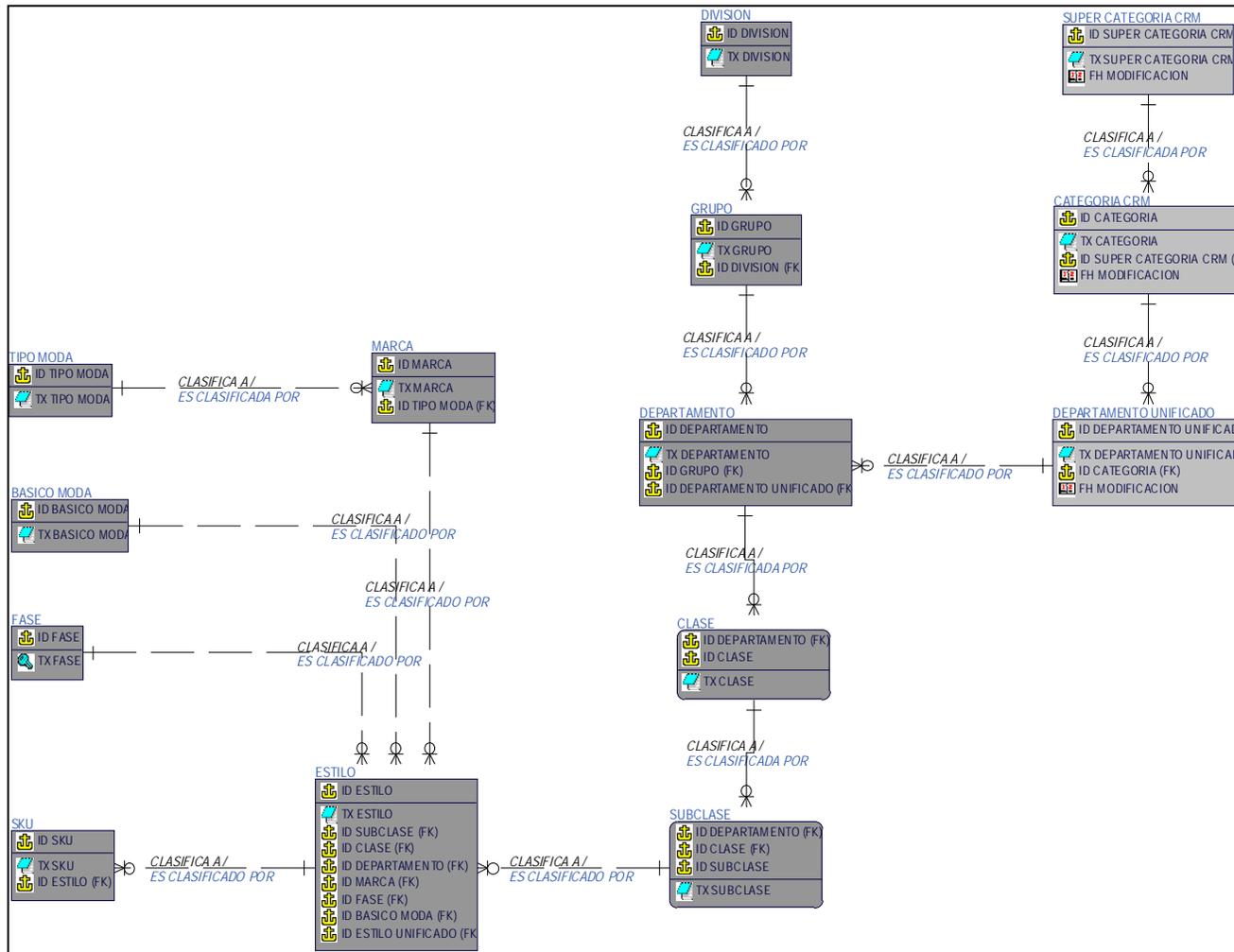
Nota. El rosa está representado con blanco y el azul con gris.



Modelo ER Data Mart CRM, área temática GRP

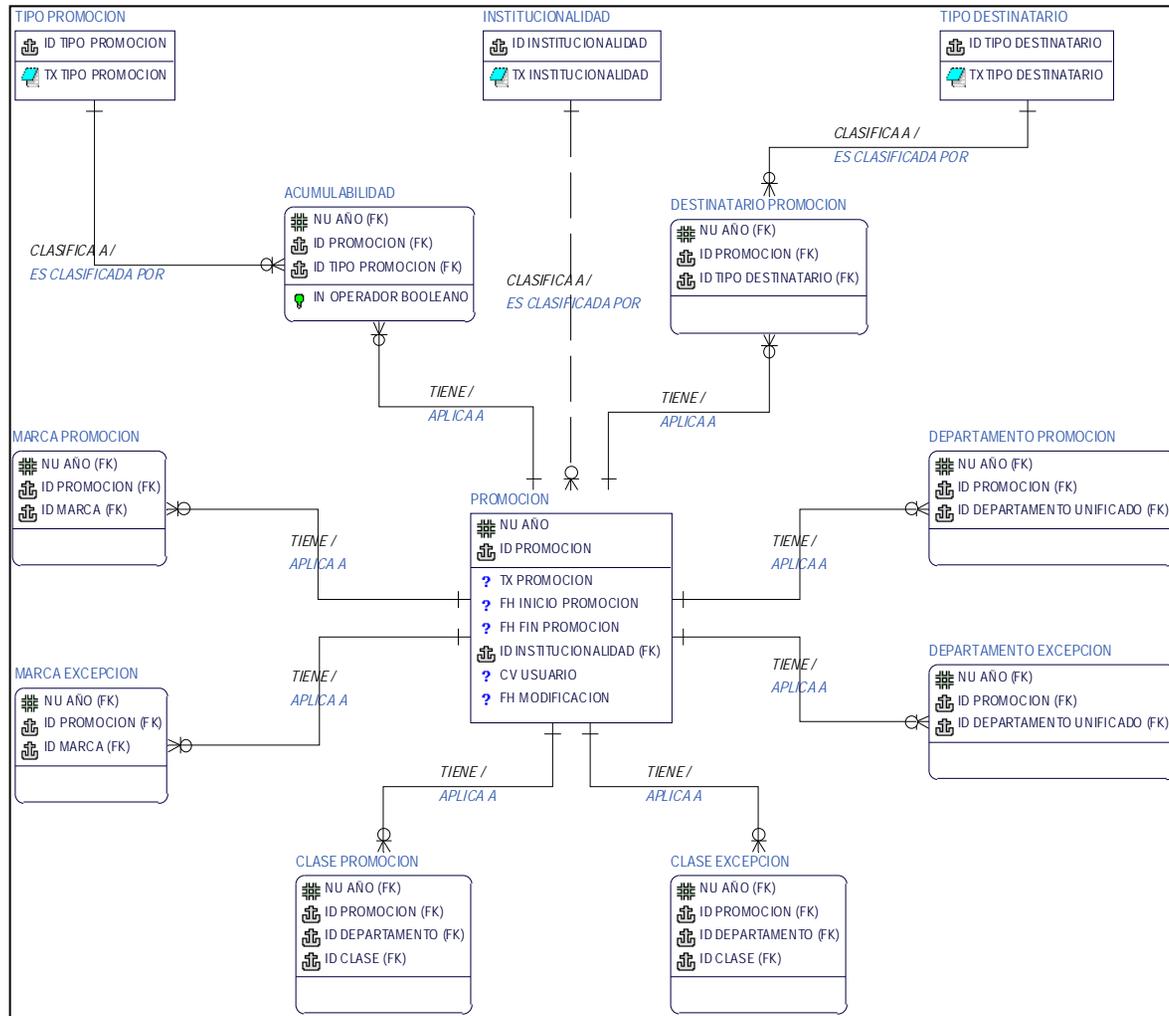
Figura A2

Nota. El rosa está representado con blanco, el azul con gris claro y el verde con gris oscuro.



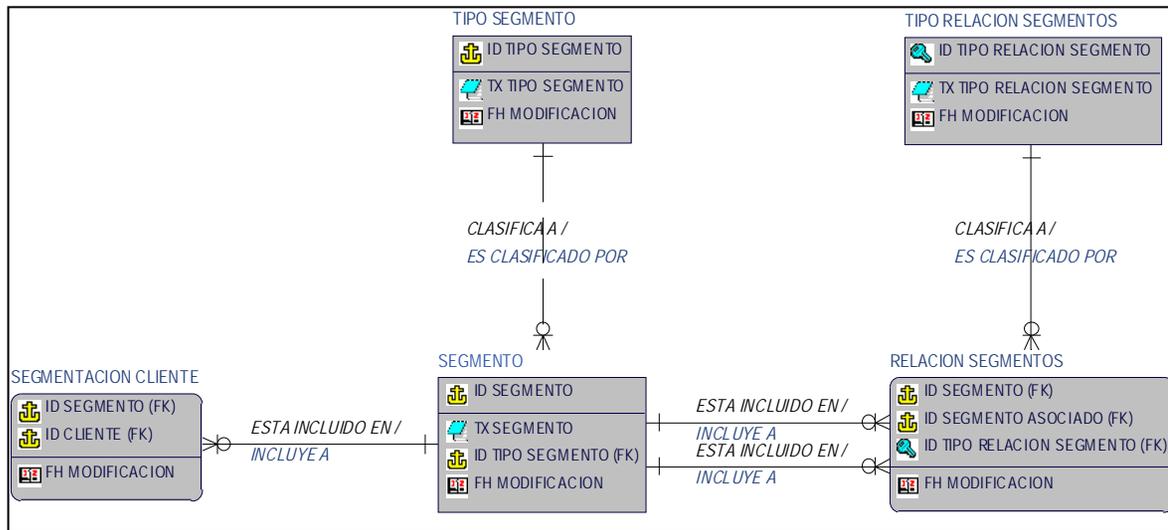
Modelo ER Data Mart CRM, área temática Artículos
Figura A3

Nota. El azul está representado con gris claro y el verde con gris oscuro.



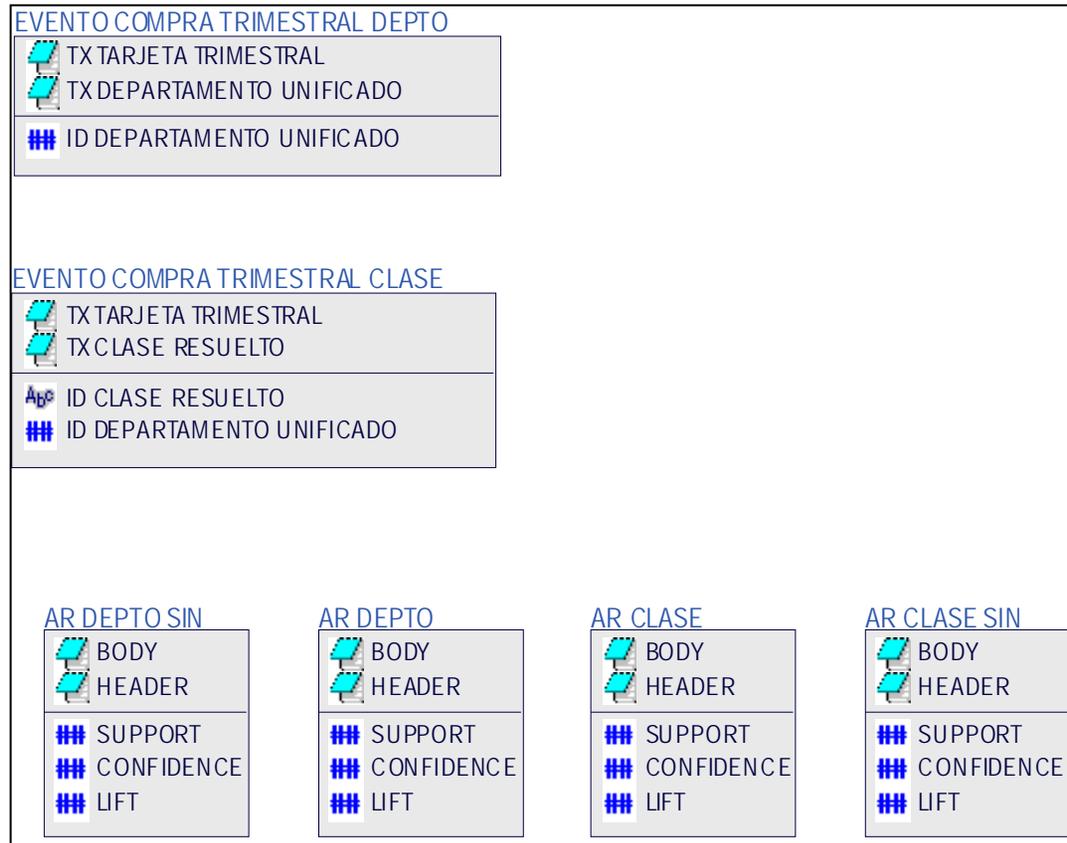
Modelo ER Data Mart CRM, área temática Promoción
Figura A4

Nota. El amarillo está representado con blanco.



Modelo ER Data Mart CRM, área temática Segmentación
Figura A5

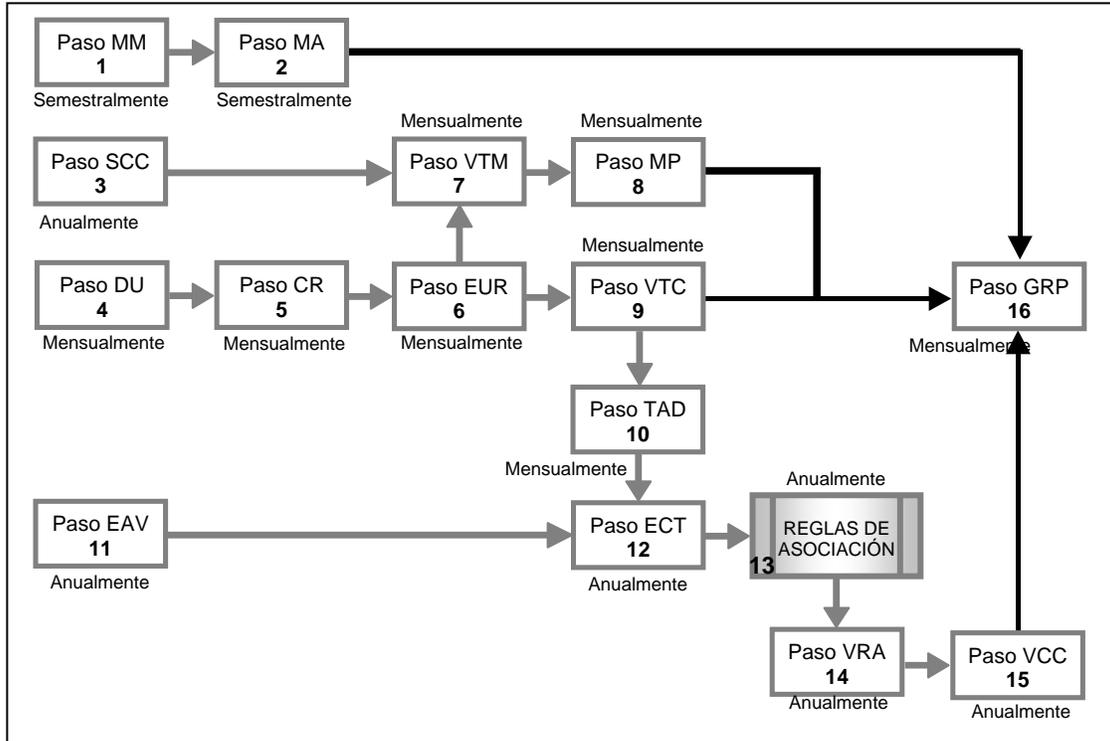
Nota. El azul está representado con gris claro.



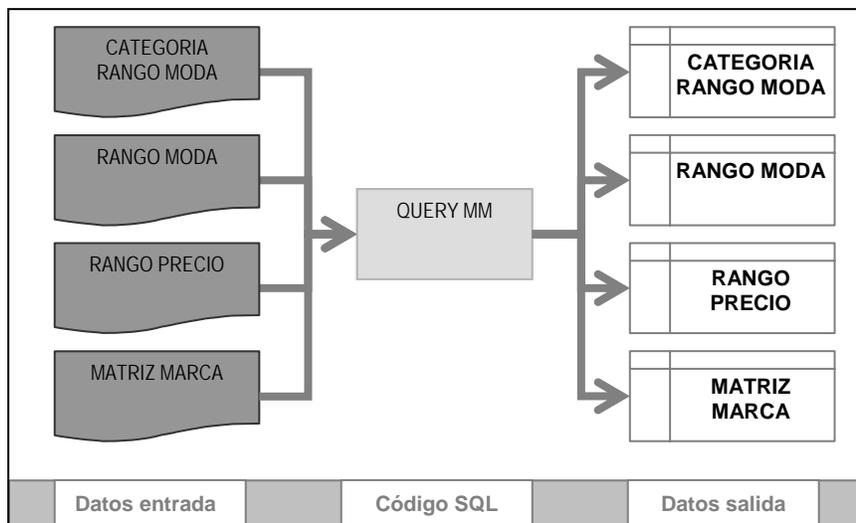
Modelo ER Data Mart CRM, Estructuras de minado
Figura A7

Anexo B

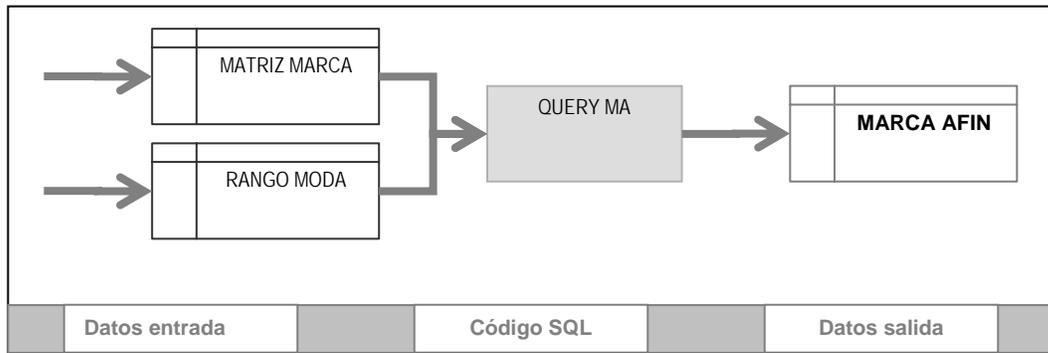
Proceso Generación de Recomendaciones Personalizadas.



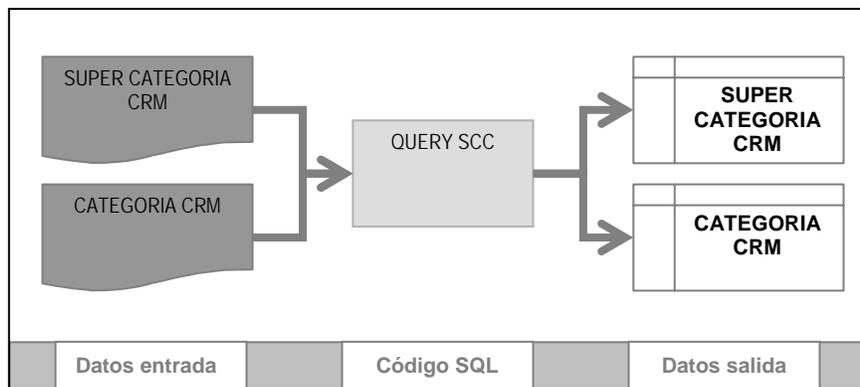
Proceso Generación de Recomendaciones Personalizadas
Figura B1



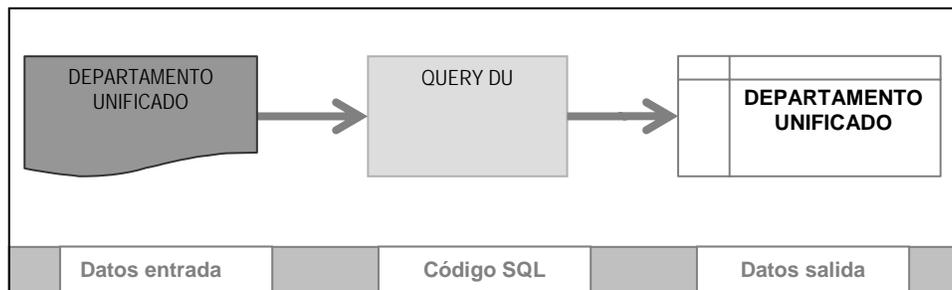
Paso 1: Matriz de marcas
Figura B2



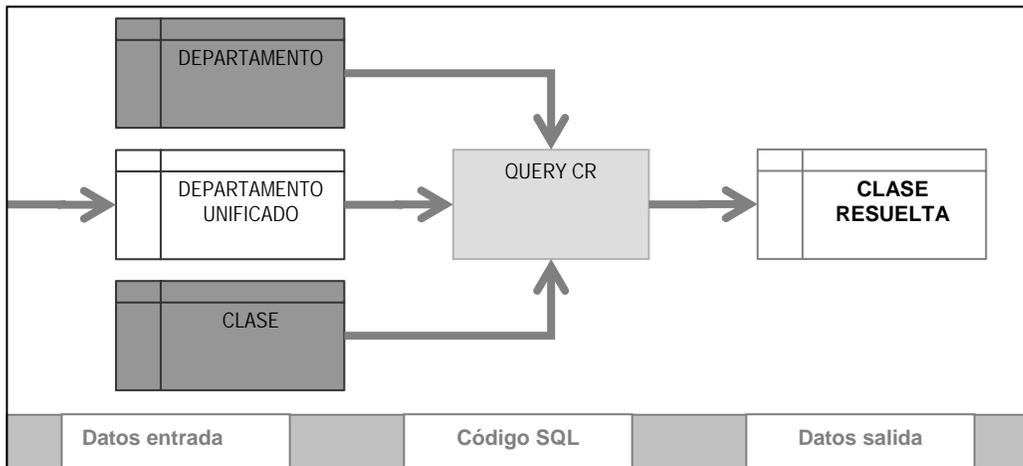
Paso 2: Marca afin
Figura B3



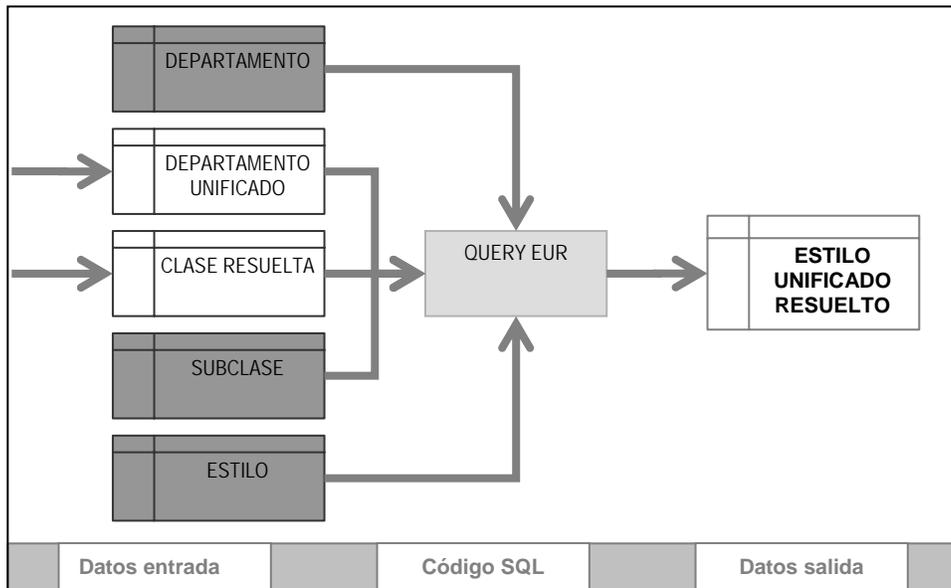
Paso 3. Súper categoría y categoría
Figura B4



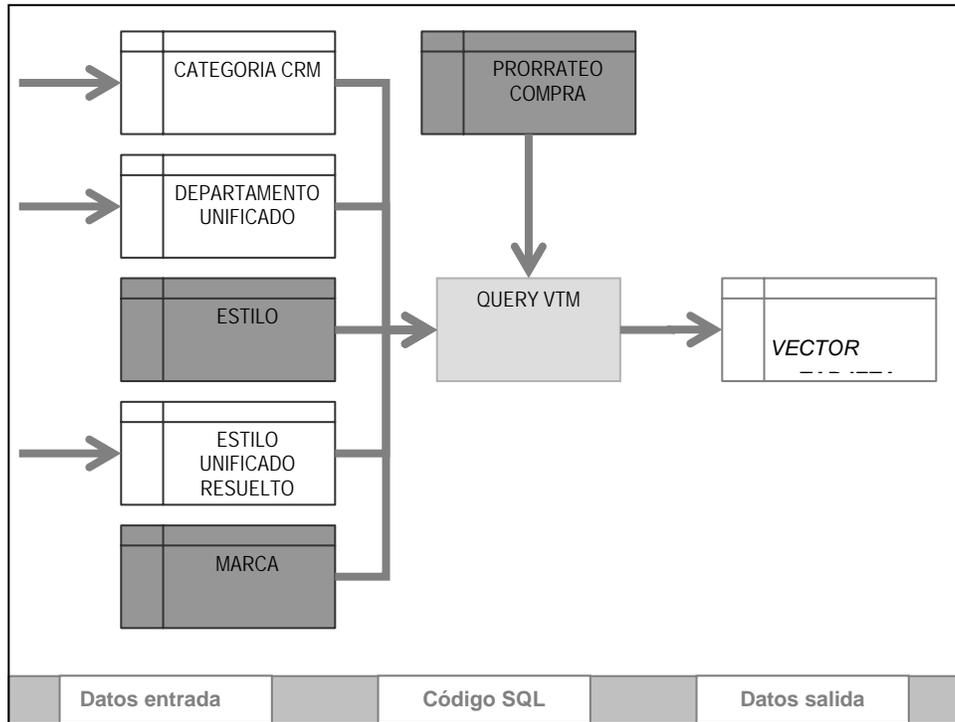
Paso 4. Departamento unificado
Figura B5



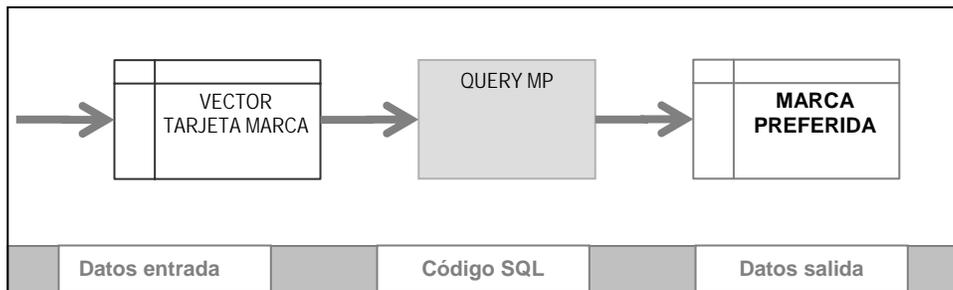
Paso 5. Clase resuelta
Figura B6



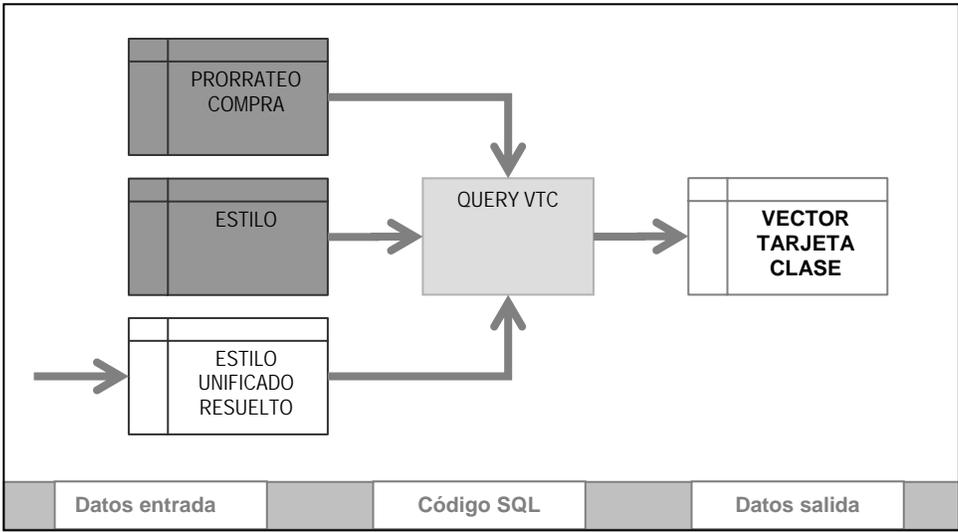
Paso 6. Estilo unificado resuelto
Figura B7



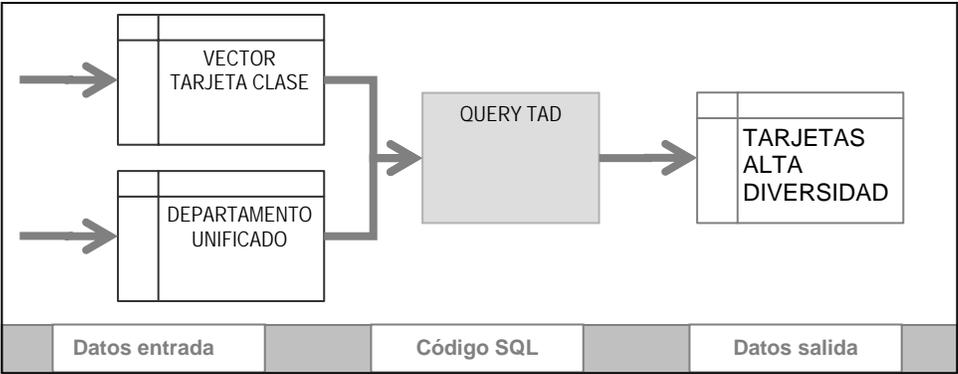
Paso 7. Vector tarjeta marca
Figura B8



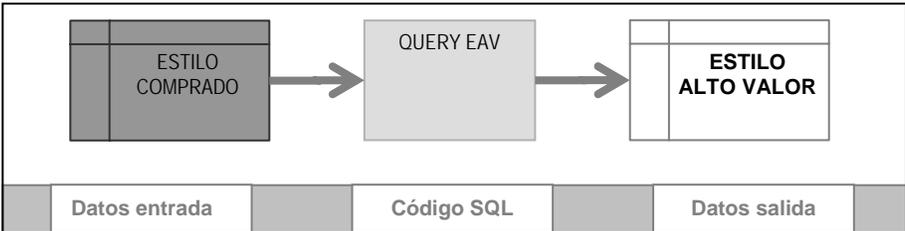
Paso 8. Marca preferida
Figura B9



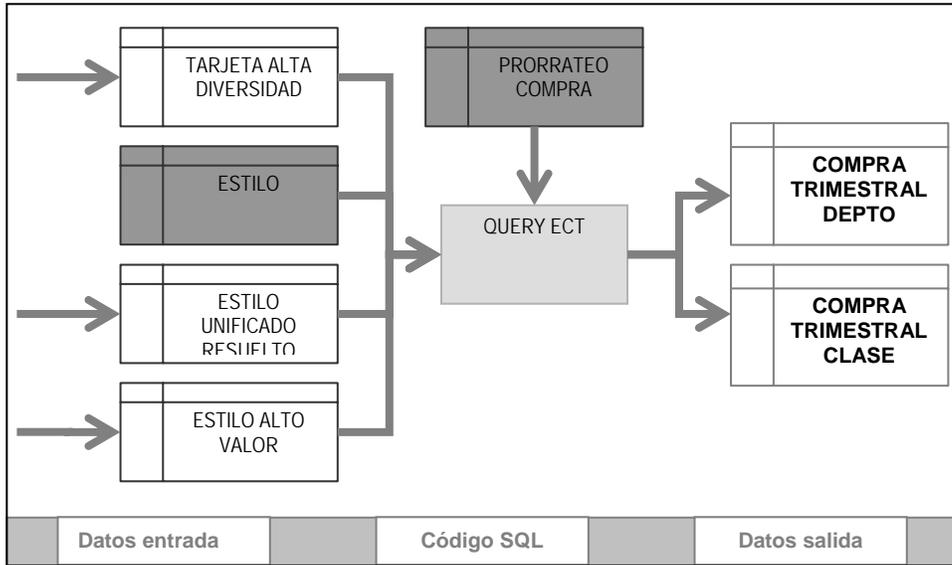
Paso 9. Vector tarjeta clase
Figura B10



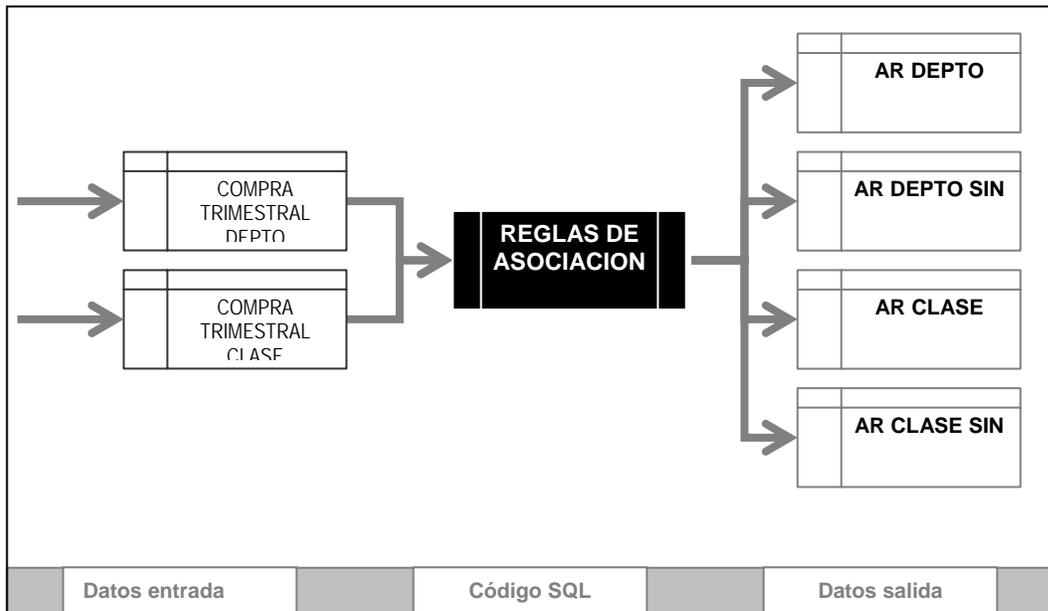
Paso 10. Tarjetas de alta diversidad
Figura B11



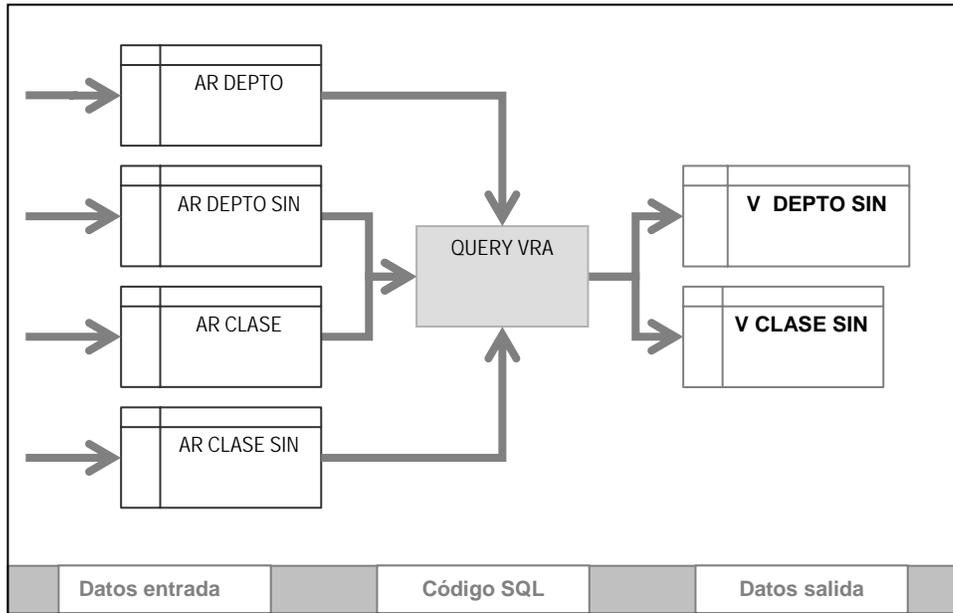
Paso 11. Estilos alto valor
Figura B12



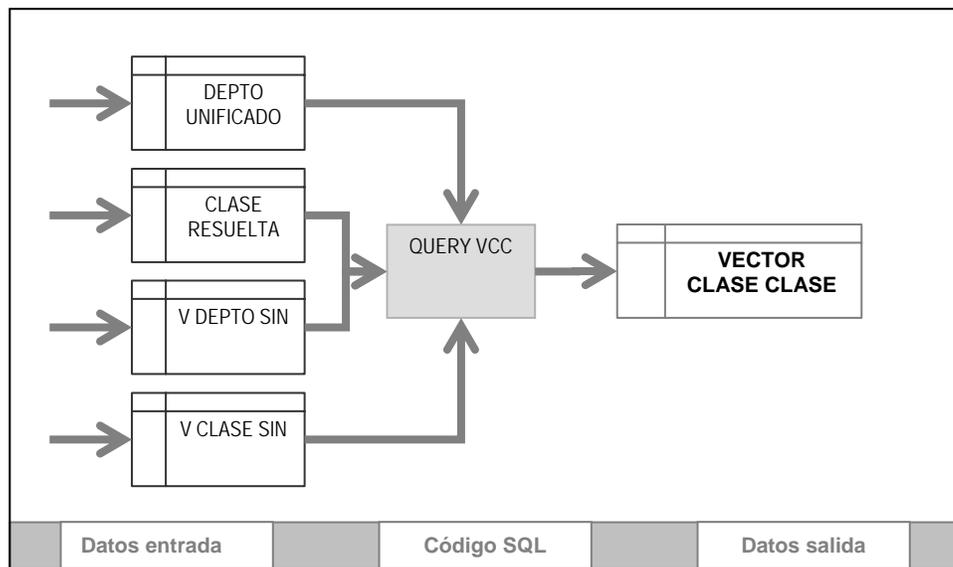
Paso 12. Eventos de compra trimestrales
Figura B13



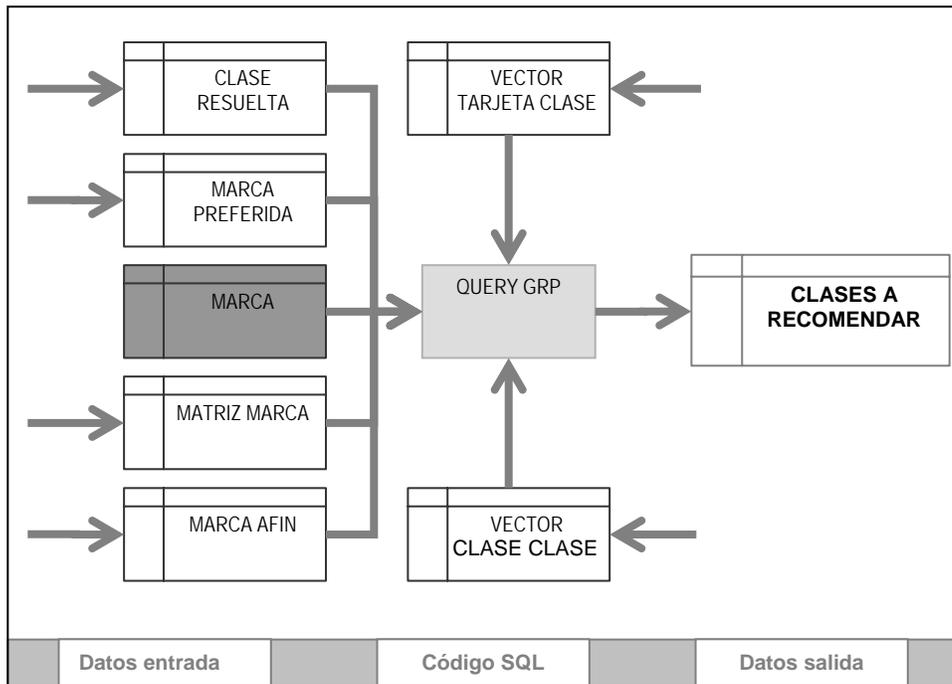
Paso 13. Reglas de asociación
Figura B14



Paso 14. Vistas de reglas de asociación
Figura B15



Paso 15. Vector clase clase
Figura B16



Paso 16. Generación de recomendaciones personalizadas
 Figura B17

Procesos en SQL.

- **Paso1 MM - Matriz de marcas.**

```
INSERT INTO MATRIZ_MARCA VALUES (13,27,7,6);
INSERT INTO MATRIZ_MARCA VALUES (23,10,11,3);
INSERT INTO MATRIZ_MARCA VALUES (23,29,11,3);
.
.
.
INSERT INTO RANGO_MODA VALUES (17,'TEMÁTICO D',7);
```

- **Paso2 MA - Marca afín.**

```
INSERT INTO MARCA_AFIN
SELECT X.ID_MARCA
, Y.ID_MARCA AS ID_MARCA_AFIN
, Y.ID_DEPARTAMENTO_UNIFICADO AS ID_DEPARTAMENTO_UNIFICADO_AFIN
, X.ID_RANGO_PRECIO
, X.ID_CATEGORIA_RANGO_MODA
FROM (SELECT A.ID_MARCA
, MIN(A.ID_RANGO_PRECIO) AS ID_RANGO_PRECIO
, MAX(B.ID_CATEGORIA_RANGO_MODA) AS ID_CATEGORIA_RANGO_MODA
FROM MATRIZ_MARCA A
, RANGO_MODA B
WHERE A.ID_RANGO_MODA = B.ID_RANGO_MODA
GROUP BY A.ID_MARCA) X
,(SELECT A.ID_MARCA, A.ID_DEPARTAMENTO_UNIFICADO
, MIN(A.ID_RANGO_PRECIO) AS ID_RANGO_PRECIO
, MAX(B.ID_CATEGORIA_RANGO_MODA) AS ID_CATEGORIA_RANGO_MODA
FROM MATRIZ_MARCA A
, RANGO_MODA B
WHERE A.ID_RANGO_MODA = B.ID_RANGO_MODA
GROUP BY A.ID_MARCA, A.ID_DEPARTAMENTO_UNIFICADO) Y
WHERE X.ID_RANGO_PRECIO = Y.ID_RANGO_PRECIO
AND X.ID_CATEGORIA_RANGO_MODA = Y.ID_CATEGORIA_RANGO_MODA
AND X.ID_MARCA <> Y.ID_MARCA
```

- **Paso3 SCC - Categorías CRM**

```
INSERT INTO CATEGORIA_CRM VALUES (2,'ACCESORIOS',1);
INSERT INTO CATEGORIA_CRM VALUES (14,'ACCESORIOS HOGAR',6);
INSERT INTO CATEGORIA_CRM VALUES (8,'AREA JOVEN CABALLERO',2);
.
.
.
INSERT INTO CATEGORIA_CRM VALUES (31,'COMPUTO',17);
```

- **Paso4 DU - Departamento unificado**

```
INSERT INTO DEPARTAMENTO_UNIFICADO VALUES (1,'ACCESORIOS DAMA',2);
INSERT INTO DEPARTAMENTO_UNIFICADO VALUES (2,'BOUTIQUE DAMA',1);
INSERT INTO DEPARTAMENTO_UNIFICADO VALUES (3,'DAMA',1);
.
.
.
INSERT INTO DEPARTAMENTO_UNIFICADO VALUES (901,'PROMOCIONES ESPEC.',29);
```

- **Paso5 CR - Clase resuelta**

```
INSERT INTO CLASE_RESUELTA
SELECT X.ID_DEPARTAMENTO_UNIFICADO
, X.ID_CLASE
, MAX(X.ID_CLASE_RESUELTA) AS ID_CLASE_RESUELTA
, MAX(X.TX_CLASE) AS TX_CLASE
, MAX(X.TX_MERCANCIA) AS TX_MERCANCIA
```

```

FROM (SELECT B.ID_DEPARTAMENTO_UNIFICADO
      , A.ID_CLASE
      , DIGITS(B.ID_DEPARTAMENTO_UNIFICADO)||' '|| DIGITS(A.ID_CLASE) AS ID_CLASE_RESUELTA
      , A.TX_CLASE
      , C.TX_DEPARTAMENTO_UNIFICADO||' '||A.TX_CLASE as TX_MERCANCIA
FROM CLASE A
      , DEPARTAMENTO B
      , DEPARTAMENTO_UNIFICADO C
WHERE A.ID_DEPARTAMENTO = B.ID_DEPARTAMENTO
      AND B.ID_DEPARTAMENTO_UNIFICADO = C.ID_DEPARTAMENTO_UNIFICADO
) X
GROUP BY X.ID_DEPARTAMENTO_UNIFICADO, X.ID_CLASE;

```

- **Paso6 EUR - Estilo unificado resuelto**

```

INSERT INTO ESTILO_UNIFICADO_RESUELTO
SELECT A.ID_ESTILO
      , A.ID_SUBCLASE
      , A.ID_CLASE
      , E.ID_DEPARTAMENTO_UNIFICADO
      , A.TX_ESTILO
      , B.TX_SUBCLASE
      , C.TX_CLASE
      , E.TX_DEPARTAMENTO_UNIFICADO
FROM ESTILO A
      , SUBCLASE B
      , CLASE_RESUELTA C
      , DEPARTAMENTO D
      , DEPARTAMENTO_UNIFICADO E
WHERE A.ID_SUBCLASE = B.ID_SUBCLASE
      AND A.ID_CLASE = B.ID_CLASE
      AND A.ID_DEPARTAMENTO = B.ID_DEPARTAMENTO
      AND A.ID_CLASE = C.ID_CLASE
      AND D.ID_DEPARTAMENTO_UNIFICADO = C.ID_DEPARTAMENTO_UNIFICADO
      AND A.ID_DEPARTAMENTO = D.ID_DEPARTAMENTO
      AND D.ID_DEPARTAMENTO_UNIFICADO = E.ID_DEPARTAMENTO_UNIFICADO;

```

```

UPDATE ESTILO_UNIFICADO_RESUELTO C
SET TX_CLASE = (SELECT MAX(A.TX_CLASE) AS TX_CLASE
               FROM CLASE A
                , DEPARTAMENTO B
               WHERE A.ID_DEPARTAMENTO = B.ID_DEPARTAMENTO
                    AND C.ID_DEPARTAMENTO_UNIFICADO = B.ID_DEPARTAMENTO_UNIFICADO
                    AND C.ID_CLASE = A.ID_CLASE
               GROUP BY B.ID_DEPARTAMENTO_UNIFICADO, A.ID_CLASE);

```

- **Paso7 VTM - Vector tarjeta marca**

```

INSERT INTO VECTOR_TARJETA_MARCA
SELECT X.NU_TARJETA
      , X.ID_MARCA
      , DEC(X.IM_P/Y.IM_P,7,6) AS PC_GASTO
      , DEC(X.IM_P,15,3) AS IM_PAGADO
      , DEC(X.IM_L,15,3) AS IM_LISTA
      , DEC(X.IM_C,15,3) AS IM_COSTO
FROM (SELECT A.NU_TARJETA
      , B.ID_MARCA
      , DEC(COUNT(*)) AS TXN
      , DEC(SUM(A.IM_PRECIO_PAGADO_UNITARIO),18,6) AS IM_P
      , DEC(SUM(A.IM_PRECIO_LISTA_UNITARIO),18,6) AS IM_L
      , DEC(SUM(A.IM_COSTO_UNITARIO),18,6) AS IM_C
FROM PRORRATEO_COMPRA A
      , ESTILO B
      , MARCA C
      , ESTILO_UNIFICADO_RESUELTO E
      , DEPARTAMENTO_UNIFICADO F
      , CATEGORIA_CRM G
WHERE A.ID_ESTILO = B.ID_ESTILO
      AND B.ID_MARCA = C.ID_MARCA

```

```

AND B.ID_ESTILO_UNIFICADO = E.ID_ESTILO_UNIFICADO
AND E.ID_DEPARTAMENTO_UNIFICADO = F.ID_DEPARTAMENTO_UNIFICADO
AND F.ID_CATEGORIA_CRM = G.ID_CATEGORIA_CRM
AND G.ID_SUPER_CATEGORIA_CRM IN (1,2,3,4)
GROUP BY A.NU_TARJETA, B.ID_MARCA
) X
,(SELECT A.NU_TARJETA
, DEC(COUNT(*)) AS TXN
, DEC(SUM(A.IM_PRECIO_PAGADO_UNITARIO),18,6) AS IM_P
, DEC(SUM(A.IM_PRECIO_LISTA_UNITARIO),18,6) AS IM_L
, DEC(SUM(A.IM_COSTO_UNITARIO),18,6) AS IM_C
FROM PRORRATEO_COMPRA A
, ESTILO B
, ESTILO_UNIFICADO_RESUELTO E
, DEPARTAMENTO_UNIFICADO F
, CATEGORIA_CRM G
WHERE A.ID_ESTILO = B.ID_ESTILO
AND B.ID_ESTILO_UNIFICADO = E.ID_ESTILO_UNIFICADO
AND E.ID_DEPARTAMENTO_UNIFICADO = F.ID_DEPARTAMENTO_UNIFICADO
AND F.ID_CATEGORIA_CRM = G.ID_CATEGORIA_CRM
AND G.ID_SUPER_CATEGORIA_CRM IN (1,2,3,4)
GROUP BY A.NU_TARJETA
) Y
WHERE X.NU_TARJETA = Y.NU_TARJETA
AND X.IM_P > 0;

```

- **Paso8 MP - Marca preferida**

```

INSERT INTO TARJETA_MARCA_PREFERIDA
SELECT A.NU_TARJETA
, MIN(A.ID_MARCA) AS ID_MARCA
FROM VECTOR_TARJETA_MARCA A
, (SELECT NU_TARJETA, MAX(PC_GASTO) AS PC_GASTO
FROM VECTOR_TARJETA_MARCA
GROUP BY NU_TARJETA) B
WHERE A.NU_TARJETA = B.NU_TARJETA
AND A.PC_GASTO = B.PC_GASTO
GROUP BY A.NU_TARJETA;

```

- **Paso9 VTC - Vector tarjeta clase**

```

INSERT INTO VECTOR_TARJETA_CLASE
SELECT X.NU_TARJETA
, X.ID_DEPARTAMENTO_UNIFICADO
, X.ID_CLASE
, DEC(X.IM_P/Y.IM_P,7,6) AS PC_GASTO
, DEC(X.IM_P,15,3) AS IM_PAGADO
, DEC(X.IM_L,15,3) AS IM_LISTA
, DEC(X.IM_C,15,3) AS IM_COSTO
FROM (SELECT A.NU_TARJETA
, C.ID_DEPARTAMENTO_UNIFICADO
, C.ID_CLASE
, DEC(COUNT(*)) AS TXN
, DEC(SUM(A.IM_PRECIO_PAGADO_UNITARIO),18,6) AS IM_P
, DEC(SUM(A.IM_PRECIO_LISTA_UNITARIO),18,6) AS IM_L
, DEC(SUM(A.IM_COSTO_UNITARIO),18,6) AS IM_C
FROM PRORRATEO_COMPRA A
, ESTILO B
, ESTILO_UNIFICADO_RESUELTO C
WHERE A.ID_ESTILO = B.ID_ESTILO
AND B.ID_ESTILO_UNIFICADO = C.ID_ESTILO_UNIFICADO
GROUP BY A.NU_TARJETA, C.ID_DEPARTAMENTO_UNIFICADO, C.ID_CLASE
) X
,(SELECT A.NU_TARJETA
, DEC(COUNT(*)) AS TXN
, DEC(SUM(A.IM_PRECIO_PAGADO_UNITARIO),18,6) AS IM_P
, DEC(SUM(A.IM_PRECIO_LISTA_UNITARIO),18,6) AS IM_L
, DEC(SUM(A.IM_COSTO_UNITARIO),18,6) AS IM_C
FROM PRORRATEO_COMPRA A
GROUP BY A.NU_TARJETA

```

```

)Y
WHERE X.NU_TARJETA = Y.NU_TARJETA
AND X.IM_P > 0;

```

- **Paso10 TAD - Tarjetas alta diversidad**

```

INSERT INTO TARJETAS_ALTA_DIVERSIDAD
SELECT X.NU_TARJETA
, COUNT(*) AS DIV
, MAX(X.PC) AS PC
FROM (SELECT A.NU_TARJETA
, B.ID_CATEGORIA_CRM
, DEC(SUM(A.PC_GASTO),7,6) AS PC
FROM VECTOR_TARJETA_CLASE A
, DEPARTAMENTO_UNIFICADO B
WHERE A.ID_DEPARTAMENTO_UNIFICADO = B.ID_DEPARTAMENTO_UNIFICADO
GROUP BY A.NU_TARJETA, B.ID_CATEGORIA_CRM
) X
GROUP BY X.NU_TARJETA
HAVING COUNT(*) >= 5
AND MAX(X.PC) <= 0.5;

```

- **Paso11 EAV - Estilos alto valor**

```

INSERT INTO ESTILOS_ALTO_VALOR
SELECT ID_ESTILO
, IM_AVG_AVG_SKU
, CT_TXN
FROM (SELECT ID_ESTILO
, DEC(AVG(IM_PRECIO_PAGADO_UNITARIO/CT_COMPRADA),15,3) AS IM_AVG_AVG_SKU
, DEC(COUNT(*),15,3) AS CT_TXN
FROM ESTILO_COMPRADO
GROUP BY ID_ESTILO
) X
WHERE IM_AVG_AVG_SKU >= 100;

```

- **Paso12 ECT - Eventos de compra trimestral**

```

INSERT INTO EVENTO_COMPRA_TRIMESTRAL_DEPTO
SELECT DISTINCT
C.TX_DEPARTAMENTO_UNIFICADO AS TX_DEPARTAMENTO_UNIFICADO
, A.NU_TARJETA||'/'||DIGITS(YEAR(A.FH_TRANSACCION))||'-'||DIGITS(QUARTER(A.FH_TRANSACCION)) AS
TX_TARJETA_TRIMESTRAL
, C.ID_DEPARTAMENTO_UNIFICADO
FROM PRORRATEO_COMPRA A
, ESTILO B
, ESTILO_UNIFICADO_RESUELTO C
, TARJETAS_ALTA_DIVERSIDAD D
, ESTILOS_ALTO_VALOR E
WHERE A.ID_ESTILO = B.ID_ESTILO
AND B.ID_ESTILO_UNIFICADO = C.ID_ESTILO_UNIFICADO
AND A.NU_TARJETA = D.NU_TARJETA
AND A.ID_ESTILO = E.ID_ESTILO;

```

```

INSERT INTO EVENTO_COMPRA_TRIMESTRAL_CLASE
SELECT DISTINCT
C.TX_DEPARTAMENTO_UNIFICADO||'/'||C.TX_CLASE AS TX_CLASE_RESUELTO
, A.NU_TARJETA||'/'||DIGITS(YEAR(A.FH_TRANSACCION))||'-'||DIGITS(QUARTER(A.FH_TRANSACCION)) AS
TX_TARJETA_TRIMESTRAL
, C.ID_DEPARTAMENTO_UNIFICADO
, DIGITS(C.ID_DEPARTAMENTO_UNIFICADO)||'-'||DIGITS(C.ID_CLASE) AS ID_CLASE_RESUELTO
FROM PRORRATEO_COMPRA A
, ESTILO B
, ESTILO_UNIFICADO_RESUELTO C
, TARJETAS_ALTA_DIVERSIDAD D
, ESTILOS_ALTO_VALOR E
WHERE A.ID_ESTILO = B.ID_ESTILO
AND B.ID_ESTILO_UNIFICADO = C.ID_ESTILO_UNIFICADO
AND A.NU_TARJETA = D.NU_TARJETA

```

AND A.ID_ESTILO = E.ID_ESTILO;

- **Paso13 Reglas Asociación**

```
INSERT INTO AR_DEPTO VALUES ('[ACCESORIOS DAMA]', '[DAMA]', 0.034288, 0.2769, 1.4957);
```

```
INSERT INTO AR_DEPTO SIN VALUES ('[VINOS Y LICORES]', '[ALIMENTOS GRAV. VTA]', 0.013474, 0.2543, 1.587);
```

```
INSERT INTO AR_CLASE VALUES ('[AREA JOVEN DAMA // Falda]', '[DAMA // Saco]', 0.004212, 0.2577, 7.2225);
```

```
INSERT INTO AR_CLASE SIN VALUES ('[AREA JOVEN DAMA // Blusa]', '[DAMA // Blusa]', 0.005801, 0.2921, 2.8965);
```

```
INSERT INTO AR_CLASE SIN VALUES ('[COCINA // Cubiertos]', '[COCINA // Accesorios]', 0.003381, 0.3865, 9.5676);
```

- **Paso14 VRA - Vistas de reglas de asociación**

```
CREATE VIEW V_DEPTO SIN AS
SELECT Y.BODY, Y.HEADER, Y.CONFIDENCE, MIN(Y.SUPPORT) AS SUPPORT
FROM (SELECT A.BODY, A.HEADER, A.CONFIDENCE, A.SUPPORT
      FROM AR_DEPTO A
      UNION
      SELECT A.BODY, A.HEADER, A.CONFIDENCE, A.SUPPORT
      FROM AR_DEPTO SIN A
     ) Y
GROUP BY Y.BODY, Y.HEADER, Y.CONFIDENCE;
```

```
CREATE VIEW V_CLASE SIN AS
SELECT Y.BODY, Y.HEADER, Y.CONFIDENCE, MIN(Y.SUPPORT) AS SUPPORT, Y.X
FROM (SELECT A.BODY, A.HEADER, A.CONFIDENCE, A.SUPPORT,
      CASE WHEN SUBSTR(BODY, 2, POSSTR(BODY, '/' )-2) = SUBSTR(HEADER, 2, POSSTR(HEADER, '/' )-2) THEN
      'INTRA'
      ELSE 'INTER'
      END AS X
      FROM AR_CLASE A
      UNION
      SELECT A.BODY, A.HEADER, A.CONFIDENCE, A.SUPPORT,
      CASE WHEN SUBSTR(BODY, 2, POSSTR(BODY, '/' )-2) = SUBSTR(HEADER, 2, POSSTR(HEADER, '/' )-2) THEN
      'INTRA'
      ELSE 'INTER'
      END AS X
      FROM AR_CLASE SIN A
     ) Y
GROUP BY Y.BODY, Y.HEADER, Y.CONFIDENCE, X;
```

- **Paso15 VCC - Vector clase clase**

```
INSERT INTO VECTOR_CLASE_CLASE
SELECT ID_CLASE_RESUELTA_BODY
, ID_CLASE_RESUELTA_HEADER
, MAX(X) AS X
, MAX(ID_D_U_BODY) AS ID_DEPTO_UNIFICADO_BODY
, MAX(ID_C_BODY) AS ID_CLASE_BODY
, MAX(ID_D_U_HEADER) AS ID_DEPTO_UNIFICADO_HEADER
, MAX(ID_C_HEADER) AS ID_CLASE_HEADER
FROM (SELECT A.ID_CLASE_RESUELTA AS ID_CLASE_RESUELTA_BODY
, B.ID_CLASE_RESUELTA AS ID_CLASE_RESUELTA_HEADER
, A.ID_DEPARTAMENTO_UNIFICADO AS ID_D_U_BODY
, A.ID_CLASE AS ID_C_BODY
, B.ID_DEPARTAMENTO_UNIFICADO AS ID_D_U_HEADER
, B.ID_CLASE AS ID_C_HEADER
, CASE WHEN A.ID_CLASE_RESUELTA = B.ID_CLASE_RESUELTA THEN -1.00
```

```

        ELSE 0.50
    END AS X
FROM CLASE_RESUELTA A
    , CLASE_RESUELTA B
WHERE A.ID_DEPARTAMENTO_UNIFICADO = B.ID_DEPARTAMENTO_UNIFICADO
-----
UNION ALL
-----
SELECT B.ID_CLASE_RESUELTA AS ID_CLASE_RESUELTA_BODY
    , C.ID_CLASE_RESUELTA AS ID_CLASE_RESUELTA_HEADER
    , B.ID_DEPARTAMENTO_UNIFICADO AS ID_D_U_BODY
    , B.ID_CLASE AS ID_C_BODY
    , C.ID_DEPARTAMENTO_UNIFICADO AS ID_D_U_HEADER
    , C.ID_CLASE AS ID_C_HEADER
    , CASE WHEN A.X = 'INTRA' THEN 0.75
        ELSE 1.00
    END AS X
FROM V_CLASE_SIN A
    , CLASE_RESUELTA B
    , CLASE_RESUELTA C
WHERE REPLACE(REPLACE(A.BODY,'[','),',') = B.TX_MERCANCIA
    AND REPLACE(REPLACE(A.HEADER,'[','),',') = C.TX_MERCANCIA
-----
UNION ALL
-----
SELECT D.ID_CLASE_RESUELTA AS ID_CLASE_RESUELTA_BODY
    , E.ID_CLASE_RESUELTA AS ID_CLASE_RESUELTA_HEADER
    , D.ID_DEPARTAMENTO_UNIFICADO AS ID_D_U_BODY
    , D.ID_CLASE AS ID_C_BODY
    , E.ID_DEPARTAMENTO_UNIFICADO AS ID_D_U_HEADER
    , E.ID_CLASE AS ID_C_HEADER
    , 0.25 AS X
FROM V_DEPTO_SIN A
    , DEPARTAMENTO_UNIFICADO B
    , DEPARTAMENTO_UNIFICADO C
    , CLASE_RESUELTA D
    , CLASE_RESUELTA E
WHERE REPLACE(REPLACE(A.BODY,'[','),',') = B.TX_DEPARTAMENTO_UNIFICADO
    AND REPLACE(REPLACE(A.HEADER,'[','),',') = C.TX_DEPARTAMENTO_UNIFICADO
    AND B.ID_DEPARTAMENTO_UNIFICADO = D.ID_DEPARTAMENTO_UNIFICADO
    AND C.ID_DEPARTAMENTO_UNIFICADO = E.ID_DEPARTAMENTO_UNIFICADO
) X
GROUP BY ID_CLASE_RESUELTA_BODY, ID_CLASE_RESUELTA_HEADER;

```

- **Paso16 GRP – Generación de recomendaciones personalizadas.**

```

INSERT INTO RECOMENDACIONES
SELECT W.NU_TARJETA, W.ID_DEPTO_UNIFICADO_BODY, W.ID_CLASE_BODY, W.SCORE, W.TX_MERCANCIA,
V.TX_MARCA
FROM (SELECT X.NU_TARJETA, X.ID_DEPTO_UNIFICADO_BODY, X.ID_CLASE_BODY, X.SCORE, X.TX_MERCANCIA
    , COALESCE(Y.ID_MARCA, MAX(Z.ID_MARCA_AFIN)) AS ID_MARCA_REC
FROM (SELECT A.NU_TARJETA
    , B.ID_DEPTO_UNIFICADO_BODY
    , B.ID_CLASE_BODY
    , SUM(A.PC_GASTO*B.X)/D.X AS SCORE
    , C.TX_MERCANCIA
    , E.ID_MARCA
    , RANK() OVER (ORDER BY SUM(A.PC_GASTO*B.X)/D.X DESC) AS RANKING
FROM VECTOR_TARJETA_CLASE A
    , VECTOR_CLASE_CLASE B
    , CLASE_RESUELTA C
    , (SELECT ID_DEPTO_UNIFICADO_BODY, ID_CLASE_BODY, SUM(ABS(X)) AS X
FROM VECTOR_CLASE_CLASE
GROUP BY ID_DEPTO_UNIFICADO_BODY, ID_CLASE_BODY) D
    , TARJETA_MARCA_PREFERIDA E
WHERE A.ID_DEPARTAMENTO_UNIFICADO = B.ID_DEPTO_UNIFICADO_HEADER
    AND A.ID_CLASE = B.ID_CLASE_HEADER
    AND B.ID_DEPTO_UNIFICADO_BODY = C.ID_DEPARTAMENTO_UNIFICADO
    AND B.ID_CLASE_BODY = C.ID_CLASE
    AND B.ID_DEPTO_UNIFICADO_BODY = D.ID_DEPTO_UNIFICADO_BODY

```

```

AND B.ID_CLASE_BODY = D.ID_CLASE_BODY
AND A.NU_TARJETA = E.NU_TARJETA
AND B.ID_CLASE_RESUELTA_BODY NOT IN (SELECT DIGITS(ID_DEPARTAMENTO_UNIFICADO)||' '||
DIGITS(ID_CLASE))
FROM VECTOR_TARJETA_CLASE F
WHERE F.NU_TARJETA = A.NU_TARJETA)
GROUP BY A.NU_TARJETA, B.ID_DEPTO_UNIFICADO_BODY, B.ID_CLASE_BODY, C.TX_MERCANCIA, D.X,
E.ID_MARCA
) X
LEFT OUTER JOIN
MATRIZ_MARCA Y
ON X.ID_MARCA = Y.ID_MARCA
AND X.ID_DEPTO_UNIFICADO_BODY = Y.ID_DEPARTAMENTO_UNIFICADO
LEFT OUTER JOIN
MARCA_AFIN Z
ON X.ID_MARCA = Z.ID_MARCA
AND X.ID_DEPTO_UNIFICADO_BODY = Z.ID_DEPARTAMENTO_UNIFICADO_AFIN
GROUP BY X.NU_TARJETA, X.ID_DEPTO_UNIFICADO_BODY, X.ID_CLASE_BODY, X.SCORE, X.TX_MERCANCIA,
Y.ID_MARCA
) W
LEFT OUTER JOIN
MARCA V
ON W.ID_MARCA_REC = V.ID_MARCA
ORDER BY W.SCORE DESC;

```

Bibliografía

- Adomavicius, G., 2001. Extending Recommender Systems: A Multidimensional Approach, IJCAI-01 Workshop on Intelligent Techniques for Web Personalization (ITWP'2001), Seattle, Washington, August 2001 (with A. Tuzhilin).
- Aggarwal, C-C., Wolf, J-L., Wu, K-L. and Yu, P-S., 1999. Horting hatches an egg: A new graph-theoretic approach to collaborative filtering. In KDD-99 Proceedings.
- Allen, R-B., 1990. User models: Theory, method and practice. International Journal of Man-Machine Studies, 1990.
- Balabanovic, M. and Shoham, Y. 1997. Fab: Content-Based, Collaborative Recommendation. Communications of ACM, March 1997, vol. 40, no.3.
- Baragoin C., Andersen, C-M., Bayerl, S., Bent, G., Lee, J. and Schommer, C., 2001. Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data. ibm.com/redbooks.
- Baudisch, P., 1999. Joining Collaborative and Content-based Filtering, In Interacting with Recommender Systems, Online Proceedings of the CHI '99 Workshop, 1999.
- Belkin, N. and Croft, B-W. 1992. Information Filtering and Information Retrieval: Two Sides of the Same Coin?
- Berzal, F., Cubero, J-C., Marín, N., Sánchez, D., Serrano, J-M., Vila, A., 2005. Association rule evaluation for classification purposes, CEDI 2005, I Congreso Español de Informática, Granada, September 2005.
- Borchers, A., Herlocker, J., Konstan, J. and Riedl, J. 1998. Ganging up on information overload. Computer.
- Candan, K-S., Li W-S., 2001. Discovering Web Document Associations for Web Site Summarization. In Proceedings International WWW Conference(10), Hong-Kong. <http://wwwconf.ecs.soton.ac.uk/archive/00000118/01/1140.pdf>
- Changchien, S-W. and Lu, Z-C., 2001. Mining association rules procedure to support on-line recommendation by customers and products fragmentation, Expert Systems with Applications, 2001.
- Chen, H-C. and Chen A-L-P., 2001. A music recommendation system based on music data grouping and user interests, in: Proceedings of the ACM International Conference on Information and Knowledge Management, Atlanta, GA, November 2001.
- Cokrowijoyo, H., Taniar, T. and D., 2004. A Framework for Mining Association Rules in Data Warehouses, Intelligent Data Engineering and Automated Learning – IDEAL 2004: 5th International, Conference, Exeter, UK. August 2004.
- Daly O., and Taniar, D., 2004. Exception Rules Mining Based on Negative Association Rules, Computational Science and Its Applications - ICCSA 2004: International Conference, Assisi, 2004.

- Demiriz, A., 2004. Enhancing Product Recommender Systems on Sparse Binary Data, *Data Mining and Knowledge Discovery*, Volume 9, Number 2, September 2004.
- Ding, Q., Ding, Q. and Perrizo, W., 2002. Association Rule Mining on Remotely Sensed Images Using P-trees, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 6th Pacific-Asia Conference, PAKDD 2002, Taipei, Taiwan, 2002.
- Duen-Ren, L. and Ya-Yueh, S., 2004. Integrating AHP and data mining for product recommendation based on customer lifetime value, *Information and Management archive* Volume 42, March 2005.
- Duen-Ren, L. and Ya-Yueh, S., 2005. Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences, *Journal of Systems and Software*, Elsevier Science Inc., Volume 77, Issue 2, 2005.
- Duen-Ren, L. and Ya-Yueh, S., 2005. Hybrid Recommendation Approaches: Collaborative Filtering via Valuable Content Information. *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) -Track 8- Volume 08*, 2005.
- Firestone, J-M., 1998. Architectural Evolution in Data Warehousing: The Coming of Distributed Knowledge Management Architecture (DKMA). EIS Web Site.
<http://www.dkms.com/papers/archev4ppt.pdf#search=%22Architectural%20Evolution%20in%20DataWarehousing%22>.
- Friedman, B., 1999. Informed consent and the design of recommender systems. Paper presented to the workshop on Interacting with Recommender Systems, 1999.
- Good, N., Schafer, J-B., Konstan, J-A., Borchers, A., Sarwar, B., Herlocker, J. and Riedl, J., 1999. Combining Collaborative Filtering with Personal Agents for Better Recommendations, In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 1999
- Hernández, J., Ramírez, M-J. and Ferri, C., 2005. *Introducción a la minería de datos*. Prentice Hall. Madrid, España, 2005.
- Hill, W-C., Stead, L., Rosenstein, M. and Fumas, G., 1995. Recommending and Evaluating Choices in a Virtual Community of Use. in *Proceedings of CHI'95*, Denver CO, May 1995.
- Hsu, C-N., Chung, H-H. and Huang, H-S., 2004. Mining Skewed and Sparse Transaction Data for Personalized Shopping Recommendation, *Data Mining Lessons Learned*, 2004.
- Inmon, B., 2000. Information Management: Charting the Course: ODS Types. *DM Review Magazine*. http://www.dmreview.com/article_sub.cfm?articleId=1749
- Jacobo, J., Zubcoff and Trujillo, J., 2004. Extending the UML for Designing Association Rule Mining Models for Data Warehouses, *Data Warehousing and Knowledge Discovery: 7th International Conference*, DaWaK 2005, Copenhagen, Denmark, 2005.
- Konstan, J-A., Riedl, J., Borchers, A., and Herlocker, J-L., 1998. Recommender Systems: A GroupLens Perspective. *Recommender Systems. Papers from 1998 Workshop*. Technical Report WS-98-08, AAAI Press, 1998.

- Konstan, J-A, Miller, B., Maltz, D., Herlocker, J., Gordon, L-R., and Riedl, J., 1997. GroupLens: Applying collaborative filtering to usenet news. *Communications of ACM*, 40, 1997.
- Lang, K., 1995. NewsWeeder: Learning to Filter Netnews, in: *Proceedings of the 12th International Conference on Machine Learning*, Tahoe City, CA, 1995.
- Lawrence, R-D., Almasi, G-S., Kotlyar, V., Viveros, M-S., and Duri, S-S., 2001. Personalization of Supermarket Product Recommendations. *Data Mining and Knowledge Discovery*, 5, 2001. Kluwer Academic Publishers, 2001.
- Lin, C-H., McLeod, D., 2002. Exploiting and Learning Human Temperaments for Customized Information Recommendation. *IMSA 2002, Internet and Multimedia Systems and Applications*. <http://www.actapress.com/PaperInfo.aspx?PaperID=25791>.
- Linden, G., Smith, B. and York, J., 2003. Amazon.com recommendations: item-to-item collaborative filtering, *Internet Computing IEEE* 7 (1), 2003.
- Liu B., Ma, Y., Wong, K-C. and Yu, P-S., 2003. Scoring the Data Using Association Rules, *Applied Intelligence*, Vol 18, No. 2, 2003.
- Lueg, C., 1999. Issues in Understanding Collaborative Filtering, *CHI'99 Interacting with Recommender Systems*, 1999.
- Massa, P., 2003. Trust-aware Decentralized Recommender Systems: PhD Research Proposal, Department of Information and Communication Technology-University of Trento-Italy, 2003.
- Ogasawara, M., Sugimori, H., Iida, Y. and Yoshida, K., 2005. Analysis Between Lifestyle, Family Medical History and Medical Abnormalities Using Data Mining Method – Association Rule Analysis, *Computer Science, Knowledge-Based Intelligent Information and Engineering Systems: 9th International Conference, KES 2005*.
- Ólafsson, S., 2005. Editorial: introduction to operations research and data mining. *Computers & Operations Research* 33.
- Pazzani, M-J., 1999. A Framework for Collaborative, Content-Based and Demographic Filtering. *Artificial Intelligence Review*, 1999.
- Perera, A., Denton, A. and Kotala, P. 2002. P-tree Classification of Yeast Gene Deletion Data, *SIGKDD Explorations*, Dec. 2002
- Ramakrishnan, N. and Ribbens, C., 2000. Mining and Visualizing Recommendation Spaces for Elliptic PDEs with Continuous Attributes, *ACM Transactions on Mathematical Software*, Vol. 26.
- Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J., 1994. GroupLens: an open architecture for collaborative filtering of Netnews, in: *Proceedings of the CSCW'94*, Chapel Hill, NC, 1994.
- Resnick, P. and Varian, H-R., 1997. Recommender systems. *Communications of the ACM*, 40, 1997.
- Rucker, J. and Polanco, M-J., 1997. Siteseer: personalized navigation for the Web, *Communications of the ACM* 40 (3), 1997.

- Schafer, J-B., Konstan, J-A. and Riedl, J., 2002. Meta-recommendation Systems: User-controlled Integration of Diverse Recommendations, Proceedings of the ACM Conference on Information and Knowledge Management. McLean, VA: ACM Press, 2002.
- Schafer, J-B. et al, 2005. The Application of Data-Mining to Recommender Systems, Encyclopedia of data warehousing and mining. Hershey, PA: Idea Group Reference, 2005.
- Schafer, J-B., 2005. DynamicLens: A Dynamic User-Interface for a Meta-Recommendation System. A workshop on the next stage of recommender systems research at the ACM Intelligent User Interfaces Conference. San Diego, CA: ACM Press., 2005.
- Seng Chee, S-H., Han, J., Wang K., 2001. RecTree: An Efficient Collaborative Filtering Method. Lecture Notes in Computer Science. Simon Fraser University, School of Computing Science, Burnaby, B.C., Canada.
<http://www-sal.cs.uiuc.edu/~hanj/pdf/dawak01.pdf>.
- Shahabi, C., Kashani, F-B., Chen, Y-S., McLeod, D., 2001. Yoda: An Accurate and Scalable Web-Based Recommendation System. CoopIS 2001.
<http://infolab.usc.edu/DocsDemos/COOPIS2001.pdf>
- Shardanand, U. and Maes, P., 1995. Social Information Filtering: Algorithms for Automating "Word of Mouth". in Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, Denver, CO, 1995.
- Shen, Y-D., Yang, Q., Zhang, Z., Lu, H., 2003. Mining the Customer's Up-To-Moment Preferences for E-commerce Recommendations, Advances in Knowledge Discovery and Data Mining: 7th Pacific-Asia Conference, 2003.
- Srikant, R. and Agrawal, R., 1995. Mining Generalized Association Rules. In Proc. of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland, September 1995.
- Sriphaew, K. and Theeramunkong, T. 2005. Revealing Topic-based Relationship Among Documents using Association Rule Mining, Artificial Intelligence and Applications, 2005.
- Tan, P-N. and Kumar, V., 2001. Mining Indirect Associations in Web Data, Department of Computer Science, University of Minnesota Minneapolis, Third International Workshop, San Francisco, CA, USA, 2001.
- Terveen, L-G., Hill, W-C., Amento, B., McDonald, D., Creter, J., 1997. Building Task-Specific Interfaces to High Volume Conversational Data. AT&T Labs – Research, 1997.
- Terveen, L-G., Hill, W-C., Amento, B., McDonald, D., Creter, J., 1997. PHOAKS: A System for Sharing Recommendations. Communications of the ACM 40.
- Tuzhilin, A. and Adomavicius, G., 1999. Integrating User Behavior and Collaborative Methods in Recommender Systems, In CHI'99 Workshop on Interacting with Recommender Systems, May 1999.
- Utley, C., 2000. Designing the Star Schema Database.
<http://www.ciobriefings.com/whitepapers/StarSchema.asp>.

- Veloso, A., Rocha, B., Carvalho, M. and Meira Jr., W., 2002. Real World Association Rule Mining, Computer Science Department, Advances in Databases : 19th British National Conference on Databases, BNCOD 19, Sheffield, UK, Universidade Federal de Minas Gerais Belo Horizonte, Brazil , 2002
- Witten, I-H. and Frank, E., 2000. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, 2000.
- Zhan, J., Matwin, S. and Chang, L-W., 2005. Private Mining of Association Rules, Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics, ISI 2005, Atlanta, GA, USA, May 2005.
- Data Mining Concepts. SQL Server 2005 Books Online. <http://msdn2.microsoft.com/en-us/library/ms174949.aspx>.
- Data Mining Techniques. Electronic textbook, StatSoft.
<http://www.statsoft.com/textbook/stdatmin.html>.
- SQL Server 2000 Resource Kit. Microsoft Corporation.
<http://www.microsoft.com/technet/prodtechnol/sql/2000/reskit/part5/c1761.msp>.
- Oracle DBA Guide to Data Warehousing and Star Schemas. Safari Books Online.
<http://safari.oreilly.com/0130325848/ch01>.