MINERIA DE DATOS Y SU PERSPECTIVA HACIA EL FUTURO

ING. ROBERTO CARLOS NARANJO
rnaranjo@unicauca.edu.co
GRUPO DE I + D EN TECNOLOGIAS DE LA INFORMACION
DEPARTAMENTO DE SISTEMAS
UNIVERSIDAD DEL CAUCA
POAYAN - COLOMBIA

1. Introducción

Se almacenan grandes cantidades de datos día a día en los procesos de negocios, estos datos son una fuente de información acerca del negocio, sus procesos y sus clientes. Lograr mejor competitividad en un negocio a partir de información contenida en los datos y usarla para estar en el negocio, talvez es de los retos más importantes hoy en día.

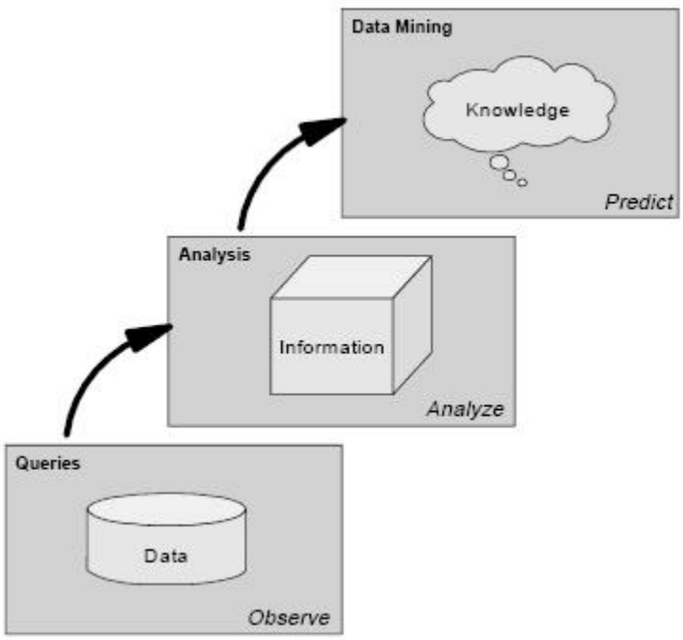

Figura 1. Evolución de consultas en DM

La evolución que ha tenido en los últimos años la minería de datos –DM [1], empieza con la ejecución de consultas contra datos operacionales resultando reportes o gráficas. El próximo paso lógico es analizar los datos resultantes con estadística tradicional o herramientas OLAP (ver sección 2), buscando tendencias o tratando de verificar una hipótesis. En el tercer paso se pueden generar modelos a partir de los datos que permitan verificar posibles circunstancias variables, estos modelos nos permiten derivar circunstancias especiales que requieren atención, en este paso, nosotros

podemos hablar de inteligencia de negocios porque se utiliza el conocimiento adquirido y se aplica al negocio.

En el primer paso se basa en preguntas, o conocimiento que un usuario ingresa y valida contra los datos disponibles. En el tercer paso es DM donde las herramientas generan conocimiento basados en los datos, este conocimiento puede ser usado para modelar su negocio sin depender de alguna suposición que no proviene de los datos del negocio.

Este artículo presenta los conceptos básicos de la DM, el proceso que debe realizar para llevar a cabo un proyecto de DM, pasando por algunas técnicas relevantes de la DM y al final presentaremos una experiencia que se realizó al interior del Grupo de Investigación en I+D en Tecnologías de la Información [20], área Tecnologías Internet, para una aplicación de comercio electrónico de tipo Business To Consumer- B2C.

2. Que es y que no es la minería de datos

La DM es el proceso de extraer información valida, útil, desconocida, y comprensible a partir de los datos y usarlos para tomar decisiones de negocios. Las características más importantes que presenta la DM son [1]:

- Proceso: DM no es una herramienta que simplemente se compra y ejecuta en un ambiente de Businness Intelligent- BI [3] y que automáticamente genera reglas para su negocio. Por el contrario tiene una serie de pasos que lo componen (ver sección 4).
- Válido: La información encontrada debe ser correcta y estadísticamente significante para soportar decisiones bien encontradas. Válido significa correctitud y completitud. Si a un gerente le interesa saber cuales son los clientes objetivo, para esto es necesario que los datos y el proceso sean válidos.
- Útil: El proceso de minería de datos debe liberar resultados que sean correctos y significantes, pero esta información debe ser útil para su negocio. Y que le permita actuar antes que sus competidores lo hagan.
- Desconocida: Debe generar nueva información. Si el proceso arroja solo información trivial esta no será de gran utilidad. Esta propiedad distingue entre verificación y descubrimiento.
- Comprensible: Los resultados del proceso de DM deben ser explicables en términos del negocio, deben generarse por ejemplo modelos que clasifican a los clientes, y la forma como se clasificaron y que factores influenciaron esta clasificación.

2.1 Olap (Online analytical process) y dataminig [4] [2]

Las herramientas de Olap permiten al usuario analizar datos de negocio rápidamente en lugar de tener que esperar mucho tiempo por los resultados de la consulta. Olap vive por el hecho que los resultados de las consultas inmediatamente generarían nuevas preguntas, las cuales deben ser procesadas rápidamente.

Con Olap usted solo encontrará información que usted buscó en primer lugar, a esto se le llama *análisis dirigido por verificación*. Los sistemas de DM encontrarán nueva información por ellos mismos, sin interferencia humana o entradas. A esto se le conoce como *análisis dirigido por*

descubrimiento. Los sistemas de DM emplean muchas técnicas para determinar relaciones claves y tendencia de los datos. Las herramientas pueden ver numerosas relaciones al mismo tiempo, resaltar las que son dominantes o excepcionales, de esta forma se gana nuevo conocimiento del negocio de sus datos existentes.

2.2 Data wharehouse [4]

Es una colección de bases de datos integradas, orientadas a temas, diseñadas para soportar las funciones de soporte a la toma de decisiones, donde cada unidad de datos es relevante en un momento del tiempo. Un Data Warehouse-DW puede ser visto como un repositorio de datos de la organización, configurada para soportar las decisiones estratégicas de mercadeo, por ejemplo. La función de un DW es almacenar datos históricos de una organización de una manera integrada que refleje las diversas facetas de la organización y el negocio.

Los datos en un DW nunca están actualizados, sino usados para responder a consultas de los usuarios finales quienes están generalmente tomando decisiones, típicamente los DW almacenan billones de registros. En muchas instancias una organización pueden tener muchos datos departamentales o locales, DW, a menudo llamados data marts. Un data mart es un DW que ha sido diseñado para reunir las necesidades de un grupo específico de usuarios, estos podrían ser grandes o pequeños dependiendo de las áreas de interés.

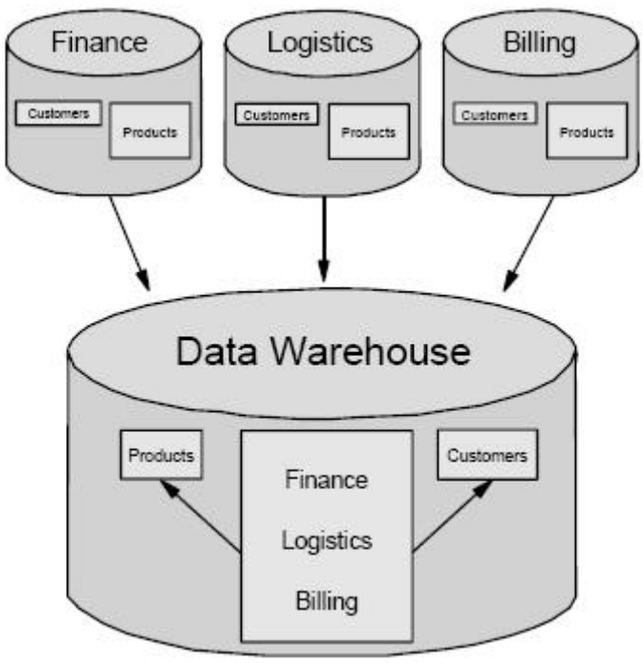

Figura 2. Data Warehouse

En la figura 2 se puede ver como un DW integra los datos de diversos sistemas fuentes en una vista corporativa de sus datos. Veamos cómo los datos cambian de ser *datos operacionales* a temas que

son importantes para su negocio, tal como productos y consumidores, que en un momento determinado servirán para ser analizados y tomar decisiones. Además de la integración de los datos, es necesario tener un sistema separado por las razones de la forma como se usan los datos:

- Las consultas sobre sistemas operacionales podrían ejecutarse contra un modelo de datos que no esta diseñado para realizar esta función, entonces se ejecutaría sobre un sistema que tiene otro propósito.
- Las consultas competirían por recursos, con la ejecución de procesos transaccionales sobre el sistema operacional, los cuales pueden causar retardos inaceptables en tiempo de ejecución.
- Los datos están constantemente cambiando, dificultando el análisis comparativo.
- La información debe ser correlacionada a través de sistemas de aplicación independientes para depurar todas las relaciones.
- Los datos operacionales son ajustados para responder rápidamente a las transacciones solicitadas, mas no para entendimiento humano.

2.3 Para hacer minería es necesario construir un DW?

No necesariamente, pero esto ayudaría mucho. La mayoría del trabajo de preparación de los datos para DM, será significativamente menor, en algunos casos eliminado. La principal ventaja de tener un DW organizado, es que se reduce el riesgo en un proyecto de DM de reunir los datos solo para una sola ejecución, ya que una de las etapas más largas de la DM es la de pre-procesamiento de los datos. De hecho una de las principales aplicaciones que tiene un DW es la DM ya que una de las principales características de los DW es la de proveer información a los usuarios para soporte a la toma de decisiones.

3. Aplicaciones y operaciones de DM

3.1 Aplicaciones de DM

Las aplicaciones de la DM dependen del ambiente del negocio en que se desee aplicar, a continuación daremos una lista posible de aplicaciones que presenta la DM (ver tabla No 1).

- Marketing directo: Es una estrategia para descubrir clientes potenciales aún desconocidos basándose en ingresos, grupo familiar, edad, etnia, afinidades y otras variables, y después localizar donde esos clientes tienden a agruparse. Preparados con los perfiles de estos clientes y una dirección para conectar a estos clientes potenciales, se diseña una intensa y focalizada campaña para transformar un significante número de clientes potenciales en reales. Una campaña típica de este tipo de marketing es el correo directo [7].

- Administración de las relaciones: Uno de los principales casos tenemos el CRM (Customers Relatioship Management), siendo este un proceso para optimizar el balance entre la inversión empresarial y la satisfacción de las necesidades de los clientes para lograr el máximo beneficio. Surge como una combinación de administración de negocios y una orientación a la tecnología como ventas asistidas por computador [7].

- Administración del canal: Consiste en el diseño de políticas y procedimientos para ganar y mantener la cooperación de varias instituciones que se encuentran del lado de las ventas de un negocio, ejemplo los proveedores [9].

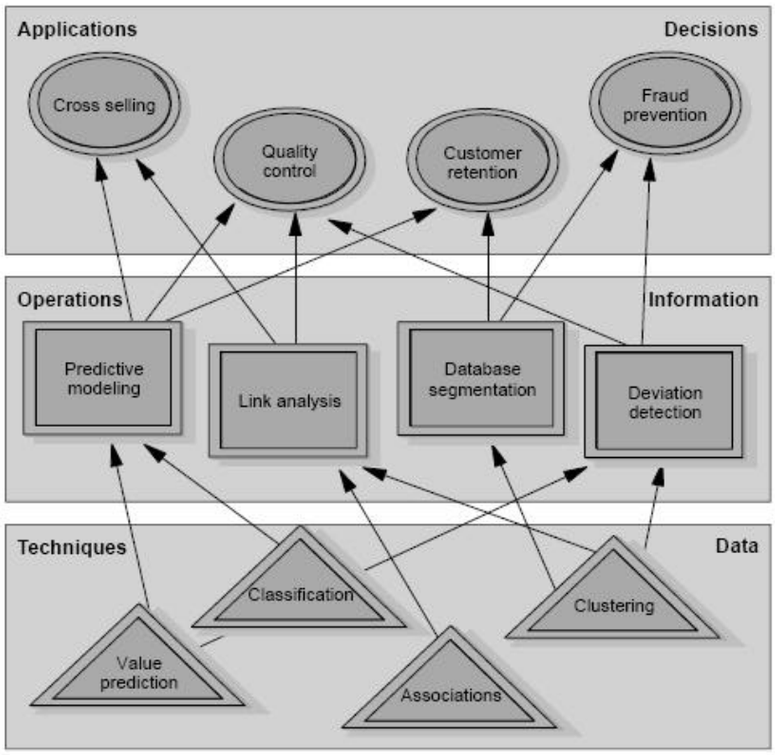

Figura 3. Aplicaciones, operaciones y técnicas [3]

- Optimización de la canasta de mercado: Consiste en optimizar las compras que realizan los clientes en cada una de las visitas para comprar utilizando los datos frecuentes de la tarjeta del comprador para seguir tendencias de compras en los grupos familiares en ciertos periodos de tiempo [10].

- Cross selling: La estrategia de impulsar nuevos productos a partir de los productos comprados anteriormente por los consumidores. Es diseñado para ampliar la confianza del cliente en la compañía y disminuir la probabilidad que el cliente migre a la competencia [11].

- Segmentación del mercado: Es una técnica que apunta a un grupo de clientes con características similares.

- Análisis de uso Web: Permite hacer un análisis de como los usuario navegan en un sitio utilizando los logs del servidor para descubrir patrones de navegación y construir mejor los sitios Web [12] .

- Forecasting: Permite predecir el comportamiento futuro del clima a partir del análisis de datos meteorológicos [13].

- Retención de clientes: Son estrategias que se implementan en una empresa para fortalecer las relaciones con sus clientes, se consideran estrategias de fidelización.

- Underwritting: Acuerdo por el cual una compañía garantiza que una emisión de acciones elevará el precio. los suscriptores se comprometen a suscribirse por cualquiera de las emisiones que no sean tomadas por el público. Ellos cobrarán una comisión por este servicio.

- Competitive analysis: Análisis de las características de las fuerzas, debilidades, y del funcionamiento de las líneas de productos de una compañía [14].

- Healthcare Fraud: Ha sido definido como medios engañosos usados por una organización para beneficiarse de acuerdos del healthcare del gobierno. Esa definición se ha ampliado más recientemente para incluir no solamente el engaño, sino también la ignorancia de las normas [15].

- Optimización del Inventario: Permite fijar los niveles de inventarios que resuelven el porcentaje de disponibilidad para los clientes, manteniendo un mínimo de inventario [16].

- Control de Calidad: Son pasos tomados para asegurar que los productos de la compañía, sean de suficientemente alta calidad [11].

- Demand Forecasting: Permite hacer un análisis de las diferentes demandas de los clientes en diferentes áreas de negocios. Además hace una estimación del panorama actual y futuro, define un marco de la decisión de uso/compra; anticipa la demanda en ciertas áreas del mercado; integra las múltiples decisiones del mercado sobre adquisición y uso, entre otros [17].

- Análisis de suscripción: Es el análisis que se realiza al porcentaje de clientes que mantienen una relación con una compañía y la terminan después de un periodo de tiempo. Por ejemplo, la suscripción de servicios de telefonía de larga distancia o revistas.

Administración de mercadeo	Administración de riesgos	Administración de procesos
Marketing directo	Predicción	Optimización del inventario
Administración de las relaciones	Fidelización del cliente	Control de calidad
Administración del canal	Análisis de suscripción	Predicción de la demanda
Optimización de la canasta de mercado	Underwritting	
Ventas cruzadas	Análisis de competencia	
Segmentación del mercado	Fraude en el sector salud	
Análisis del uso Web		

Tabla 1. Áreas de aplicación de la DM [4]

3.2 Operaciones de DM

Como visualizamos en la figura 3, las aplicaciones de la DM son soportadas por las operaciones de la DM, a continuación daremos una breve descripción de las principales operaciones:

- Modelo predictivo: Consiste en predecir el valor de un atributo usando ejemplos. Ejemplos: Asignar la categoría de riesgo a nuevos clientes o determinar la probabilidad de un cliente conteste a un correo enviado.

- Segmentación de bases de datos: Usando atributos, encontrar grupos de registros donde los registros en cada grupo tengan atributos similares, sin embargo existe diferencia entre los grupos. Ejemplo: agrupe los clientes basados en su comportamiento o también es usada como paso preparatorio para construir un modelo predictivo.

- Análisis de enlaces: Encontrar enlaces o conexiones entre registros dentro de una transacción o sobre el tiempo. Ejemplo: Analizar cuales productos se venden juntos para optimizar su disposición de venta o inventario o también utilizar este tipo de operación para analizar cuestionarios o series de tratamientos médicos.

- Detección de la desviación: Encontrar registros o series de registros, en su base de datos que contienen valores que no podría encontrar, que no cumplen con las características de los demás. Ejemplo: use esto para encontrar patrones de comportamiento fraudulento o realizar control de la calidad en sus procesos de producción.

4. El proceso de DM [1][3][4]

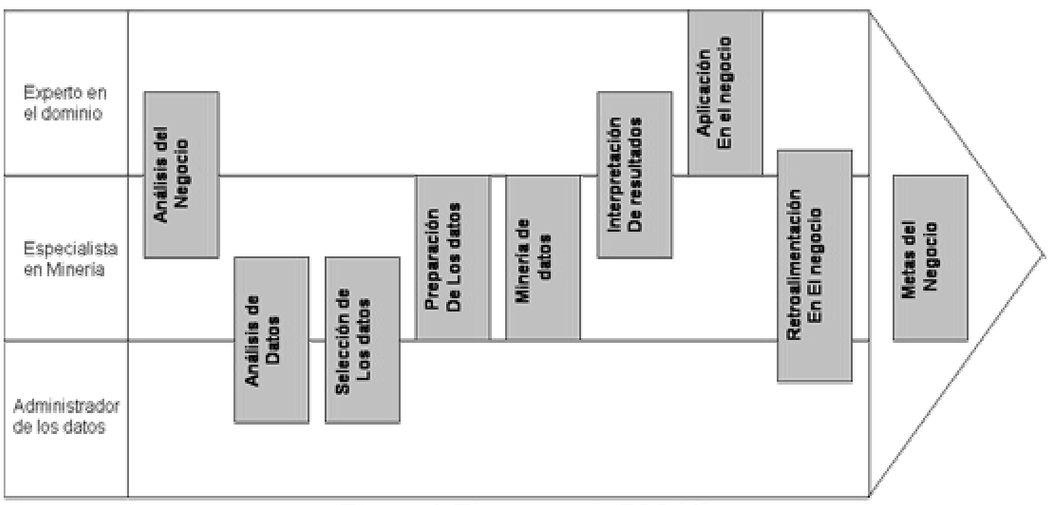

Figura 4. El proceso de DM [4]

La figura 4 muestra los pasos del proceso de DM, no se asume que sea un proceso lineal sino orientado por iteraciones, estas iteraciones configuran un resultado final configurado al negocio. A continuación daremos una breve descripción de cada una de las etapas de dicho proceso.

- Análisis del negocio: Generalmente los negocios tiene una misión que se refleja en los objetivos y sus estrategias. Para ayudarle a cumplir dichas metas podría utilizarse la DM, por ejemplo: incrementar la satisfacción de los clientes, disminuir el fraude, optimización del inventario, entre otros. Sería el caso de una empresa que desea incrementar la satisfacción de los clientes, se podría especificar este requerimiento como: "Ganar comprensión en los factores que influencian la satisfacción de los clientes, de tal forma que nos permita influenciar estos factores"; lo que permitiría cumplir con este requerimiento de incrementar la satisfacción de los clientes.

  Como se aprecia en la gráfica en esta etapa es necesaria la intervención del experto del dominio que se concentra en especificar los requerimientos del negocio, y el especialista de DM guarda la factibilidad de estos requerimientos desde el punto de vista de DM y especifica las operaciones de DM necesarias para satisfacer los requerimientos.

- Análisis de datos: El segundo paso es encontrar como los requerimientos son representados en los datos. Esta conexión es realizada a través de las operaciones de DM definida en el paso anterior, porque el especialista de DM sabe que clase de datos son necesarios para ejecutar estas operaciones. Una de las metas en este punto es identificar las fuentes de datos disponibles y extraer los datos que se necesitan para el análisis preliminar en la preparación para la minería. Esta etapa involucra el especialista de DM quien realiza la mayoría de las tareas y el administrador de las bases de datos quien dará soporte al acceso de los datos.

- Selección de los datos: Es un paso de construcción del "Data Mart", especialmente para el proyecto de DM. Este data mart podría ser virtual, quiere decir, que provee una vista de los datos de un DW, o una copia de los datos de un DW. Durante esta etapa los datos son limpiados e integrados con datos de otras fuentes de acuerdo al análisis de los datos. También utiliza muestreo para reducir el tamaño de la población. En esta etapa involucra el especialista de DM. El especialista utiliza la salida de la etapa anterior, para especificar cuales datos son necesarios y recoge los datos especificando las consultas para ejecutarlas en la etapa siguiente o enseguida para construir una base de datos separada.

- Preparación de los datos: Cuando los datos están disponibles para la minería, estos usualmente necesitan algún tipo de preparación antes de realizar la minería. La preparación de esos datos son evaluados en la etapa de análisis de los datos. En esta etapa se realizan las siguientes tareas:

  - ✓ Verificación la calidad de los datos: Los datos generalmente presentan valores llamados *outliers*, estos están fuera del rango normal que se espera. Si los datos aún son realistas, una conversión algorítmica de los datos convertiría los datos a un rango más pequeño. De otra manera, sería necesario remover cada uno de los datos que contienen esos valores, o los atributos dentro de todos los registros. Uno de los más grandes problemas de esto, son los valores perdidos, en el primer caso los registros presentan valores perdidos, en este caso no se pueden utilizar estos registros; en el segundo caso un

atributo tiene perdido ciertos valores, en este caso es mejor eliminar el atributo. Otra forma de manejar valores perdidos es la imputación, que se trata de adivinar el valor perdido para prevenir descartar registros, se realiza de una de las siguientes formas:

1. Ponga un valor randomico de los otros registros.
2. Tome la moda, mediana o media del atributo de los otros registros.
3. Construya un modelo estadístico de los valores de los otros registros y seleccione un valor randomico de acuerdo a esta distribución.
4. Trate de predecir el valor perdido con técnicas estadísticas o minería de los valores encontrados en registros similares.

- ✓ Manipulación de los datos: Los atributos disponibles demuestran a veces un alto grado de la intercorrelación, esto significa que la misma información está presente en varios atributos. Para evitar que esta información domine demasiado, podemos utilizar varias técnicas para la reducción de la dimensión. Estas técnicas intentan reducir la cantidad de atributos al mínimo, de tal forma que se contenga la misma información. Son a veces también útiles para acelerar el proceso de minería de los datos debido al número reducido de atributos. En algunos de estos casos utilizamos Clustering, ya que divide los datos en grupos más homogéneos, ahora si estos grupos son diferentes, será mejor manejar modelos para cada uno de ellos. La última acción es dividir el conjunto de datos en un conjunto de entrenamiento y uno de prueba. Esto nos sirve ya que los datos de prueba nos permite validar el modelo, esto previene el *overfitting*, que consiste en que el modelo esta completamente adoptado para un conjunto de entrenamiento y no es lo suficientemente general para manejar otros datos fuera del conjunto.

- Minería de datos: Este paso consiste en ejecutar las técnicas de minería de datos contra los datos, y puede involucrar la ejecución de varias técnicas combinadas, o la ejecución de una técnica cuyos resultados sirvan como entrada para la ejecución de otra técnica. En este caso utilizamos los conjuntos de entrenamiento y de prueba definidos en la preparación de los datos para validar el modelo construido. Si por alguna razón la DM no provee los resultados adecuados, en este caso tendríamos que repetir los pasos anteriores. En el peor de los casos la conclusión es que los datos disponibles no permiten construir un modelo adecuado.

- Interpretación de resultados: Estos dependen fuertemente de los datos visualizados del paso de DM. Las mismas gráficas que usamos para evaluar la calidad del modelo son ahora usadas para explicar los resultados en términos del negocio.

- Aplicación en el negocio: No solo basta con que los resultados sean lógicos, sino que también puedan verificarse en el ambiente de negocio. Si por ejemplo se detecto un alto segmento de ganancia dentro de su total de clientes, se debe decidir como será la estrategia para acercarse a esos clientes y al mismo tiempo, y como medir si ellos realmente representarán ganancias como lo predice el modelo. La toma de decisiones será soportada por el experto del dominio.

5. Técnicas de DM [1][2]

A continuación presentaremos algunas de las técnicas más importantes utilizadas en la DM. Tenga en cuenta que nos son las únicas actualmente existentes.

5.1 Reglas de asociación

Uno de los casos típicos de uso de esta técnica es el *análisis de la Canasta de Mercado,* esta analiza los hábitos de compra de los clientes, encontrando asociaciones entre los diferentes artículos que el consumidor coloca en su canasta de compras. El descubrir tales asociaciones puede ayudar a los minoristas, a desarrollar estrategias para saber cuales productos son comprados juntos por los clientes. Tal información podría ayudar a colocar los productos de manera apropiada en los estantes del supermercado de tal forma que los productos se vendan mejor.

Cuales grupos de artículos son probablemente comprados por los clientes en una visita a un supermercado? Para esto se requiere un análisis de canasta de mercado, donde se analizarán los datos de las transacciones de las compras del supermercado. El resultado será talvez un plan de mercadeo, estrategias de publicidad, promociones de ciertos productos, o también el diseño de un catálogo de productos. Un caso de esto podría ser diseñar un esquema diferente de la disposición de los artículos en la tienda.

Sea $J = \{I_1, I_2, \ldots, I_m\}$ un conjunto de productos. Sea D un conjunto de transacciones de base de datos donde cada transacción T es un conjunto de productos tal que T *subconjunto de* J. Sea A un conjunto de productos, donde A transacción T es decir contiene A si y solo si A *subconjunto de* T. Una regla de asociación es una implicación de la forma A *entonces* B, donde A *subconjunto propio de* J y B *subconjunto propio de* J, y $A \cap B = \emptyset$. La regla A *entonces* B cumple en el conjunto de transacciones D con un soporte s, donde s es el porcentaje de transacciones en D que contienen A U B, esto es la *P(AUB)*. La regla A *entonces* B tiene una confianza c, en el conjunto de transacciones D, si c es el porcentaje de transacciones en el conjunto D conteniendo A tal que contiene B. Esta es la probabilidad condicional *P(A|B). Las reglas que se llaman fuertes se consideran aquellas que cumplen con un mínimo de soporte y un mínimo de confianza*.

5.2 Clasificación

Esta técnica predice etiquetas de clase. Tales modelos nos sirven por ejemplo para categorizar préstamos en un banco, de tal forma que concluya que un préstamo es riesgoso o seguro realizarlo. Dentro de las aplicaciones más importantes tenemos la aprobación de créditos bancarios, diagnóstico médico, selección del mercado, entre otros.

El proceso de clasificación de los datos, se desarrolla en dos pasos así (Ver figura 5):

- En el primer paso se construye un modelo describiendo un conjunto de clases de datos o conceptos, el modelo es construido analizando tuplas de bases de datos descritas en los atributos. Cada tupla se asume que pertenece a una clase predefinida, determinada por un atributo llamado etiqueta de clase. Las tuplas de datos analizadas para construir el modelo se llama *conjunto de datos de entrenamiento*. A este tipo de aprendizaje se le llama aprendizaje supervisado, ya que las etiquetas de clases son conocidas.

- En el segundo paso se estima la exactitud predictiva del modelo, para esto utilizamos el método *holdout method* que emplea un conjunto de datos de prueba con clases etiquetadas. Las muestras son randómicamente seleccionadas y son independientes de la muestra de entrenamiento. La exactitud del modelo sobre los datos de prueba es el porcentaje que fue correctamente clasificado. Si la exactitud del modelo es aceptable, el modelo puede ser usado para clasificar tuplas de datos en el futuro, para las cuales las etiquetas de clases son desconocidas.

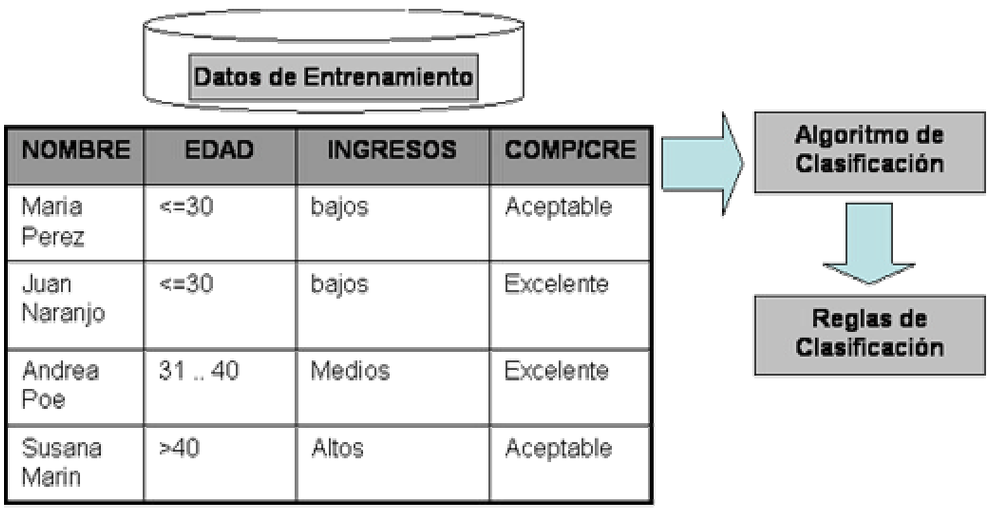

Figura 5. Proceso de Clasificación

Entre las técnicas de clasificación más importante encontramos: clasificación por árboles de decisión, clasificación bayesiana, clasificación por backpropagation (utiliza algoritmos de redes neuronales), clasificación basada en conceptos de minería de reglas de asociación, entre otros.

5.3 Análisis de Clustering

Clustering es el proceso por el cual se agrupan los datos en clases o clusters y los objetos dentro de un cluster tiene alto grado de similaridad, en comparación con otros. Pero son diferentes con respecto a otros objetos que se encuentran en otros clusters. Esta técnica determina las diferencias entre los valores de los atributos que describen a los objetos, con frecuencia para realiza esto, se utiliza la medida de la distancia. Dentro de las aplicaciones del clustering más importantes tenemos el descubrimiento de distintos grupos entre los clientes, basados en sus patrones de compra. En biología puede ser usado para encontrar taxonomías de plantas y animales, categorizar genes con funcionalidades similares, entre otras aplicaciones.

Existen diferentes tipos de algoritmos para clustering, la selección depende del tipo de datos que tengamos y del propósito de la aplicación. Los métodos más utilizados para clustering son: método de particiones, método jerárquico, método basado en la densidad, método basado en Grid, método basado en modelamiento. Para ilustrar como funciona alguno de ellos hablaremos a continuación del primero.

Dada una base de datos de n objetos y k número de clusters para formar, el algoritmo de partición organiza los objetos en k particiones (donde k <= n), donde cada partición representa un cluster. Los clusters se forman para optimizar un criterio de partición objetivo, a menudo llamado función semejanza, tal como la distancia, así que los objetos dentro del cluster son similares, y los objetos de distintos clusters son diferentes, en términos de los atributos de la base de datos. Uno de los algoritmos empleados en esta técnica es el k-means (ver figura 6).

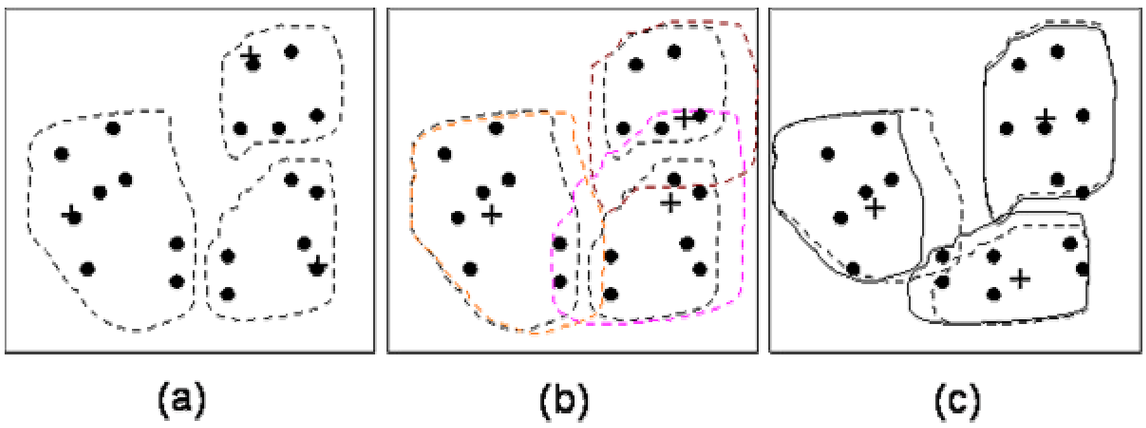

Figura. 6. Clustering utilizando el método k-means

Tenemos un conjunto de objetos localizados en el espacio de la figura 6(a), siendo k = 3, esto quiere decir que el número de clusters es de tres. Seleccionamos tres objetos arbitrariamente como centros iniciales del clusters, donde cada centro lo representan un "+", cada objeto es distribuido a cada cluster dependiendo del centro del cluster que sea más cercano, este se aprecia en la figura 6(a). Cada valor medio en cada cluster es recalculado de acuerdo a los objetos en cada uno. De acuerdo a estos nuevos centros los objetos son redistribuidos en los clusters basados en el centro más cercano, otra vez. Esta nueva redistribución se analiza en la figura 6(b). Este proceso es iterativo y se repetirá hasta cuando ya no se realiza ninguna redistribución en ningún cluster, obteniendo la figura 6(c), siendo los clusters finales los que se representan en líneas continuas.

6. Estudio de un caso para Comercio Electrónico - B2C

El proyecto "Comunidad Virtual de Negocios para el Cauca – Plataforma Comercial en Internet - CVN" [19], fue desarrollado con el apoyo de Colciencias [18], la Universidad del Cauca y la Empresa Cygnus Tecnología. Este proyecto adecuó las Tecnologías de la Información y las Comunicaciones -TIC a las necesidades comerciales específicas del departamento del Cauca-Colombia. Tres aspectos del problema se destacan: ausencia de asociatividad-colaboración entre empresas, distanciamiento cliente/empresa por causa del deterioro de los canales de comunicación y desconocimiento sobre las ventajas del uso de las TIC. CVN planteó un concepto de hacer negocios que congrega las empresas de la región, mediante un ambiente virtual que ofrece servicios de valor agregado orientados a la PUBLICIDAD, B2B y B2C, como alternativa de solución al problema planteado. La CVN, ofrece condiciones culturales y tecnológicas que motivan la asociatividad y

colaboración entre empresas, así como la formulación de estrategias de promoción de productos y servicios que favorezcan el acercamiento eficaz de estas con sus clientes [6].

Específicamente para el módulo de B2C que desarrolló la CVN se presentó un servicio de Inteligencia de negocios con el fin de identificar reglas que nos permitieran conocer como serían los hábitos de compra de los clientes de la CVN para implementar estrategias orientadas a los mismos, haciendo uso de la plataforma virtual. Este ofrece el análisis al Administrador de la CVN, basado en el flujo de transacciones entre empresas y clientes, determinado a partir de los pedidos hechos, para que desde un análisis de esas transacciones se puedan determinar la probabilidad de que un producto puede ser comprado, a partir de otros productos relacionados. Para implementar este módulo se empleó la *técnica de Reglas de asociación utilizando el algoritmo apriori [2]*. Para el desarrollo de este servicio (FRAMEWORK DE MINERIA) se crea una tabla de transacciones a partir de las tablas de Productos, Pedidos y DetallesPedido (ver diseño de la base de datos figura 8), residentes en la base de datos. Esta tabla de transacciones es creada por el FrameWork de Minería de datos, el cual se encarga de ejecutar el algoritmo Apriori sobre esta información para enviar los resultados del análisis para ser interpretado por el Administrador de la CVN. El algoritmo Apriori básicamente encuentra la cuenta de soporte y confianza de registros eliminando los registros que no reúnen un mínimo soporte y confianza desde una lista final de reglas creadas.

En las figura 7 se aprecia el análisis de casos de uso para este servicio, en este diagrama se aprecia que uno de los principales casos de uso es el de análisis de información, que es el encargado de realizar esta labor en conjunto con el administrador de la CVN.

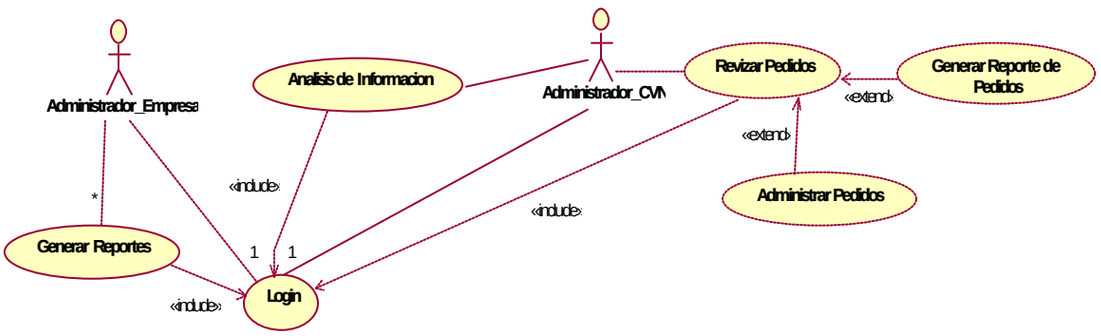

Figura. 7. Diagrama de casos de uso del servicio de Minería de datos [5]

En la figura 8 se aprecia el diseño de la base de datos para el servicio de B2C que nos servirá para alimentar el módulo de Minería de datos, se destacan entidades tales como cliente, pedidos, productos, carro de compras, empresas, que soportarán las transacciones que se realicen entre los clientes y las empresas pertenecientes a la CVN.

Por último una vez el FrameWork fue adaptado a la CVN, se desarrollaron algunas pruebas con datos de prueba para validad las reglas encontradas en la base de datos. Ver figura 9.

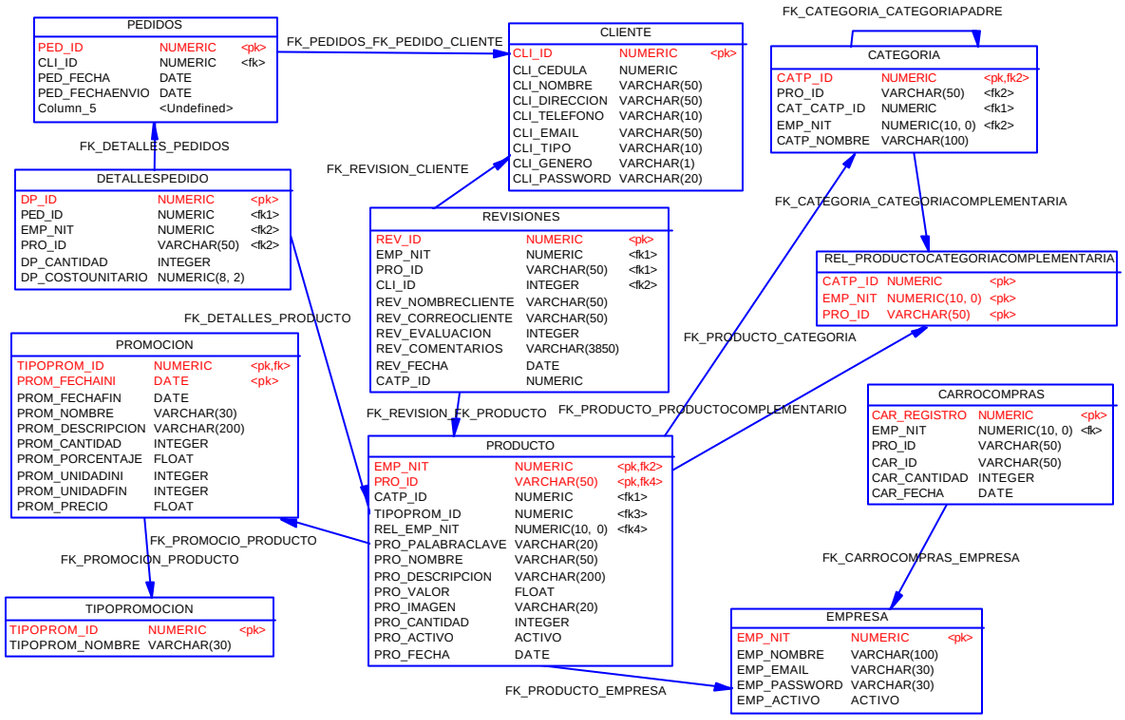

Figura. 8. Diseño de la base de datos para el módulo B2C [5]

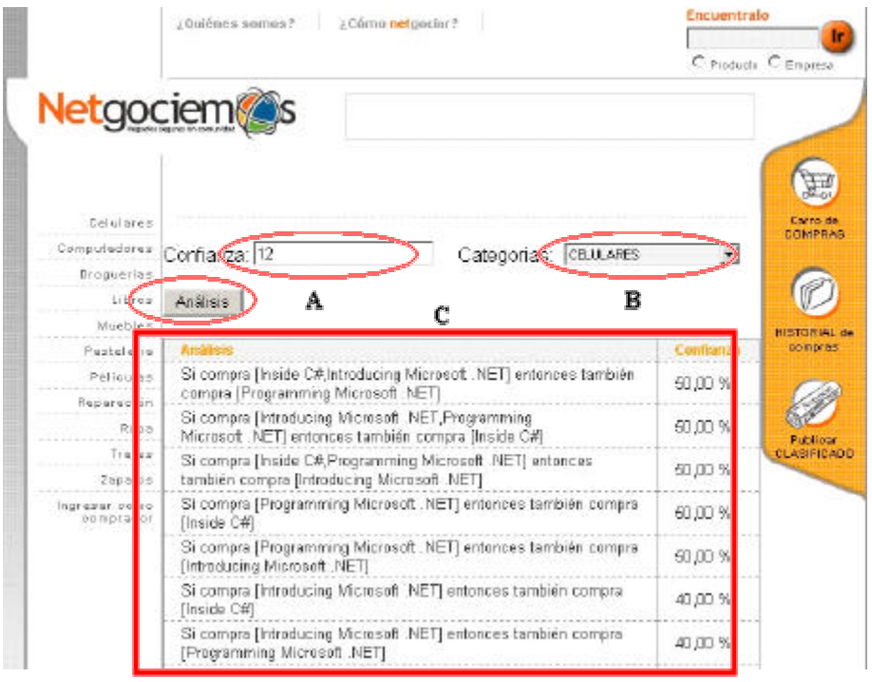

Figura. 9. Screen Shot del Servicio de Análisis de Información [5]

7. Conclusiones

- La DM es un proceso sistemático e iterativo que provee los pasos necesarios para desarrollar adecuadamente un proyecto de minería.
- Uno de los pasos más importantes en el desarrollo de un proyecto de minería es la selección de los datos, además que es un paso clave ya que depende directamente de esto el resultado final.

- Conocer bien el negocio al que se quiere implantar un proceso de minería es muy importante ya que esto nos permitirá seleccionar que operación de minería y técnica emplearemos.

- Actualmente las diferentes aplicaciones que tiene la DM son variadas y nos permite como lo vimos en el caso de estudio de la CVN, a los negocios de hoy en día fortalecer su competitividad, mediante la adaptación de estas tecnologías.

- Para el departamento del Cauca existen muchas posibilidades de aplicar este tipo de tecnologías ya que hay gran cantidad de empresas pequeñas y medianas en diferentes sectores, que están creciendo y mirarían con buenos ojos este tipo de opciones para ganar competitividad a nivel regional y nacional.

- La Universidad del Cauca como motor de investigación y desarrollo debe seguir apuntando a desarrollar este tipo de aplicaciones de DM ya que son muy costosas de adquirir y sería muy ventajoso contar con aplicaciones propias al alcance del mercado.

Bibliografía

[1] Discovering Data Mining From Concepts To Implementation. Meter Cabena, Pablo Hadjinian, Rolf Stader, Jaap Verhees, Alessandro Zanasi. Prentice Hall. 1997. USA.

[2] Data Mining Concepts and Techniques. Jiawei Han, Micheline Kamber. Morgan Kaufmann Publisher. 2001. USA.

[3] Enhance Your Business Applications. Corinne Baragoin, Ronnie Chan, Helena Gottschalk, Gregor Meyer, Paulo Pereira, Jaap Verhees. IBM. 2002. USA.

[4] Intelligent Miner for Data: Enhance Your Business Intelligence. Joerg Reinschmidt, Helena Gottschalk, Hosung Kim, Damiaan Zwietering. IBM. 1999. USA.

[5] Módulo de Business To Consumer- B2C para la Comunidad Virtual de Negocios para el Departamento del cauca-CVN". Roberto Naranjo, José Luis Dorado, Andrés Ortiz. Universidad del Cauca. 2003. Popayán-Colombia.

[6] Building a Virtual E-commerce Community. Jorge Moreno, Roberto Naranjo, Luz Marina Sierra, Martha Mendoza. Proceedings of IADIS Internacional Conference. 2004. Lisboa-Portugal.

[7] http://www.businessforum.com/target2.html. Visitada Marzo 2006.

[8] https://extranet.iwi.unisg.ch/public/ckp_web.nsf/SysWebRessources/crm_km/$FILE/f262.pdf. Visitada Marzo 2006.

[9] http://www.bitpipe.com/tlist/. Visitada Marzo 2006.

[10] http://www.msa.com/ims/publishing/. Visitada Marzo 2006.

[11] http://www.investorwords.com/. Visitada Marzo 2006.

[12] http://iit.demokritos.gr/~paliourg/papers/SMC99.pdf. Visitada Marzo 2006.

[13] http://www.thefreedictionary.com/. Visitada Marzo 2006.

[14] http://www.moneyglossary.com. Visitada Marzo 2006.

[15] http://www.quitamfyi.com. Visitada Marzo 2006.

[16] http://www.johngalt.com. Visitada Marzo 2006.

[17] http://www.nationalanalysts.com/marketing. Visitada Marzo 2006.

[18] http://www.colciencias.gov.co. Visitado Marzo 2006.

[19] http://www.netgociemos.com. Visitado Marzo 2006.

[20] http://gti.unicauca.edu.co. Visitado Marzo 2006.