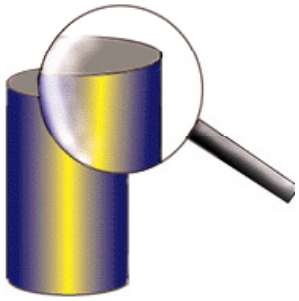
Universidad Nacional del Nordeste

Facultad de Ciencias Exactas, Naturales y Agrimensura

Trabajo de Adscripción

Minería de Datos



Adscripta: Sofia J. Vallejos - L.U.: 37.032

Materia: Diseño y Administración de Datos

Director: Mgter. David Luis la Red Martínez

Licenciatura en Sistemas de Información
Corrientes - Argentina

2006

Índice General

Índice de Figuras

Capítulo 1

Inteligencia de Negocios

1.1 Business Intelligence

Algo peor que no tener información disponible es tener mucha información y no saber qué hacer con ella. La Inteligencia de Negocios o Business Intelligence (BI) es la solución a ese problema, pues por medio de dicha información puede generar escenarios, pronósticos y reportes que apoyen a la toma de decisiones, lo que se traduce en una ventaja competitiva. La clave para BI es la información y uno de sus mayores beneficios es la posibilidad de utilizarla en la toma de decisiones. En la actualidad hay una gran variedad de software de BI con aplicaciones similares que pueden ser utilizados en las diferentes áreas de la empresa, tales como, ventas, marketing, finanzas, etc. Son muchas las empresas que se han beneficiado por la implementación de una sistema de BI, además se pronostica que con el tiempo se convertirá en una necesidad de toda empresa.

En este nuevo mundo, la información reina afirma Geoffrey A. Moore, Director de Chasm Group. Vivimos en una época en que la información es la clave para obtener una ventaja competitiva en el mundo de los negocios. Para mantenerse competitiva una empresa, los gerentes y tomadores de decisiones requieren de un acceso rápido y fácil a información útil y valiosa de la empresa. Una forma de solucionar este problema es por medio del uso de Business Intelligence o Inteligencia de Negocios.

1.2 ¿Qué es Business Intelligence?

La Inteligencia de Negocios o Business Intelligence (BI) se puede definir como el proceso de analizar los bienes o datos acumulados en la empresa y extraer una cierta inteligencia o conocimiento de ellos. Dentro de la categoría de bienes se incluyen las bases de datos de clientes, información de la cadena de suministro, ventas personales y cualquier actividad de marketing o fuente de información relevante para la empresa.

BI apoya a los tomadores de decisiones con la información correcta, en el momento y lugar correcto, lo que les permite tomar mejores decisiones de negocios. La información adecuada en el lugar y momento adecuado incrementa efectividad de cualquier empresa.

La tecnología de BI no es nueva, ha estado presente de varias formas por lo menos en los últimos 20 años, comenzando por generadores de reportes y sistemas de información ejecutiva en los 80's Afirma Candice Goodwin . Entiéndase como sinónimos de tecnología de BI los términos aplicaciones, soluciones o software de inteligencia de negocios.

Para comprender mejor el concepto se sita el siguiente ejemplo. Una franquicia de hoteles a nivel nacional que utiliza aplicaciones de BI para llevar un registro estadístico del porcentaje promedio de ocupación del hotel, así como los días promedio de estancia de cada huésped, considerando las diferencias entre temporadas. Con esta información ellos pueden:

- Calcular la rentabilidad de cada hotel en cada temporada del año.
- Determinar quién es su segmento de mercado.
- Calcular la participación de mercado de la franquicia y de cada hotel.
- Identificar oportunidades y amenazas.

Estas son sólo algunas de las formas en que una empresa u organización se puede beneficiar por la implementación de software de BI, hay una gran variedad de aplicaciones o software que brindan a la empresa la habilidad de analizar de una forma rápida por qué pasan las cosas y enfocarse a patrones y amenazas. En la figura 1.1 de la página 3 se ilustra lo antes mencionado.

Figura 1.1: Ilustración del B.I.

1.3 Componentes de Business Intelligence

Todas las soluciones de BI tienen funciones parecidas, pero deben de reunir al menos los siguientes componentes:

- Multidimensionalidad: la información multidimensional se puede encontrar en hojas de cálculo, bases de datos, etc. Una herramienta de BI debe de ser capaz de reunir información dispersa en toda la empresa e incluso en diferentes fuentes para así proporcionar a los departamentos la accesibilidad, poder y flexibilidad que necesitan para analizar la información. Por ejemplo, un pronóstico de ventas de un nuevo producto en varias regiones no está completo si no se toma en cuenta también el comportamiento histórico de las ventas de cada región y la forma en que la introducción de nuevos productos se ha desarrollado en cada región en cuestión.

- Data Mining: Las empresas suelen generar grandes cantidades de información sobre sus procesos productivos, desempeño operacional, mercados y clientes. Pero el éxito de los negocios depende por lo general de la habilidad para ver nuevas tendencias o cambios en las tendencias. Las aplicaciones de data mining pueden identificar tendencias y comportamientos, no sólo para extraer información, sino también para descubrir las relaciones en bases de datos que pueden identificar comportamientos

que no muy evidentes.

- Agentes: Los agentes son programas que piensan. Ellos pueden realizar tareas a un nivel muy básico sin necesidad de intervención humana. Por ejemplo, un agente pueden realizar tares un poco complejas, como elaborar documentos, establecer diagramas de flujo, etc.

- Data Warehouse: Es la respuesta de la tecnología de información a la descentralización en la toma de decisiones. Coloca información de todas las áreas funcionales de la organización en manos de quien toma las decisiones. También proporciona herramientas para búsqueda y análisis.

Capítulo 2

Introducción

2.1 Descubrimiento de Conocimiento en Bases de Datos (KDD)

En los últimos años, ha existido un gran crecimiento en nuestras capacidades de generar y colectar datos, debido básicamente al gran poder de procesamiento de las máquinas como a su bajo costo de almacenamiento.

Sin embargo, dentro de estas enormes masas de datos existe una gran cantidad de información oculta, de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información. El descubrimiento de esta información oculta es posible gracias a la Minería de Datos (Data Mining), que entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad, pero es el descubrimiento del conocimiento (KDD, por sus siglas en inglés) que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados.

Así el valor real de los datos reside en la información que se puede extraer de ellos, información que ayude a tomar decisiones o mejorar nuestra comprensión de los fenómenos que nos rodean. Hoy, más que nunca, los métodos analíticos avanzados son el arma secreta de muchos negocios exitosos. Empleando métodos analíticos avanzados para la explotación de datos, los negocios incrementan sus ganancias, maximizan la eficiencia operativa, reducen

costos y mejoran la satisfacción del cliente

2.2 Concepto del KDD

De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación de la información y ese modelo representen un valor agregado, entonces nos referimos al conocimiento. En la figura 2.1 de la página 7 se ilustra la jerarquía que existe en una base de datos entre datos, información y conocimiento. Se observa igualmente el volumen que presenta en cada nivel y el valor que los responsables de las decisiones le dan en esa jerarquía. El área interna dentro del triángulo representa los objetivos que se han propuesto. La separación del triángulo representa la estrecha unión entre dato e información, no así entre la información y el conocimiento.

La capacidad de generar y almacenar información creció considerablemente en los últimos tiempos, se ha estimado que la cantidad de datos en el mundo almacenados en bases de datos se duplica cada 20 meses. Es así que hoy las organizaciones tienen gran cantidad de datos almacenados y organizados, pero a los cuales no les pueden analizar eficientemente en su totalidad.

Con las sentencias SQL se puede realizar un primer análisis, aproximadamente el 80% de la información se obtiene con estas técnicas. El 20% restante, que la mayoría de las veces, contiene la información más importante, requiere la utilización de técnicas más avanzadas.

El Descubrimiento de Conocimiento en Bases de Datos

(KDD) apunta a procesar automáticamente grandes cantidades de datos para encontrar conocimiento útil en ellos, de esta manera permitirá al usuario el uso de esta información valiosa para su conveniencia.

El KDD es el Proceso no trivial de identificar patrones válidos, novedosos, potencialmente

útiles y, en última instancia, comprensibles a partir de los datos .

(Fallad et al., 1996)

El objetivo fundamental del KDD es encontrar conocimiento útil, válido,

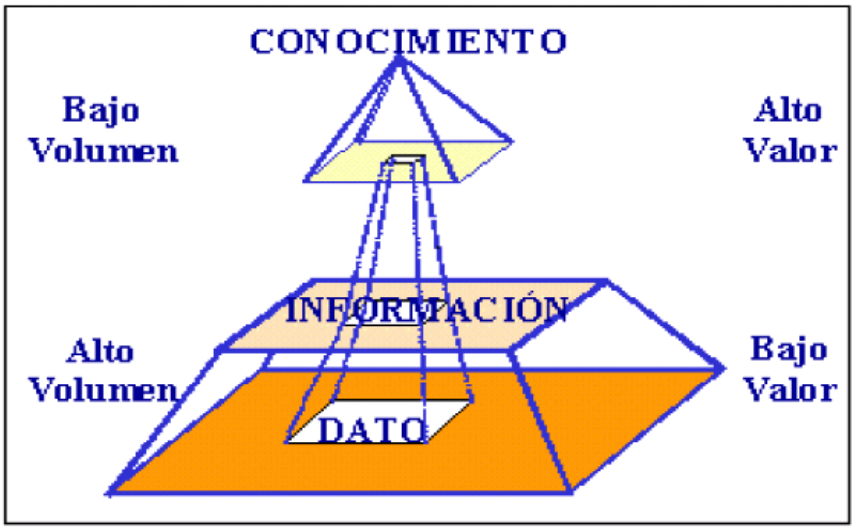


Figura 2.1: Jerarquía del Conocimiento.

relevante y nuevo sobre un fenómeno o actividad mediante algoritmos eficientes, dadas las crecientes órdenes de magnitud en los datos. Al mismo tiempo hay un profundo interés por presentar los resultados de manera visual o al menos de manera que su interpretación sea muy clara. Otro aspecto es que la interacción humano-máquina deberá ser flexible, dinámica y colaboradora. El resultado de la exploración deberá ser interesante y su calidad no debe ser afectada por mayores volúmenes de datos o por ruido en los datos. En este sentido, los algoritmos de descubrimiento de información deben ser altamente robustos.

2.2.1 Metas

Las metas del KDD son:

- Procesar automáticamente grandes cantidades de datos crudos.
- Identificar los patrones más significativos y relevantes.
- Presentarlos como conocimiento apropiado para satisfacer las metas del usuario.

2.3 Relación con otras disciplinas

- KDD nace como interfaz y se nutre de diferentes disciplinas:
- Sistemas de información / bases de datos: tecnologías de bases de datos y bodegas de datos, maneras eficientes de almacenar, accesar y manipular datos.
- Estadística, aprendizaje automático / IA (redes neuronales, lógica difusa, algoritmos genéticos, razonamiento probabilístico): desarrollo de técnicas para extraer conocimiento a partir de datos.
- Reconocimiento de patrones: desarrollo de herramientas de clasificación.
- Visualización de datos: interfaz entre humanos y datos, y entre humanos y patrones.
- Computación paralela / distribuida: cómputo de alto desempeño, mejora de desempeño de algoritmos debido asu complejidad y a la cantidad de datos.
- Interfaces de lenguaje natural a bases de datos.

2.4 El proceso de KDD

El proceso de KDD consiste en usar métodos de minería de datos(algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con preprocesamientos y post-procesamientos.En la figura2.2 de la pagina 9se ilustra el proceso de KDD.

Se estima que la extracción de patrones (minería) de los datos ocupa solo el 15% - 20% del esfuerzo total del proceso de KDD.

El proceso de descubrimiento de conocimiento en bases de datos involucra varios pasos:

- Determinar las fuentes de información: que pueden ser útiles y dónde conseguirlas.

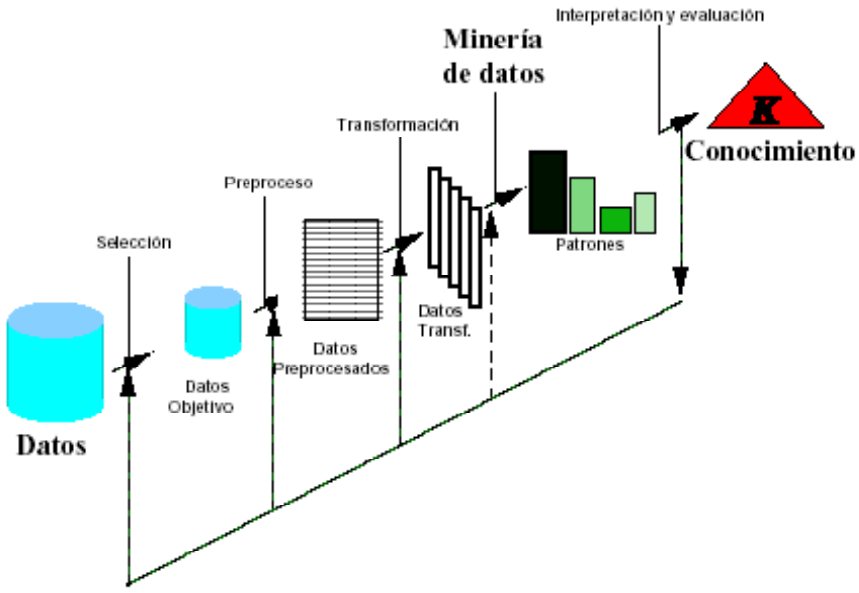


Figura 2.2: Proceso de KDD

- Diseñar el esquema de un almacén de datos (Data Warehouse): que consiga unificar de manera operativa toda la información recogida.
- Implantación del almacén de datos: que permita la *navegación* y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados. Esta es la etapa que puede llegar a consumir el mayor tiempo.
- Selección, limpieza y transformación de los datos que se van a analizar: la selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos).La limpieza y prepocesamiento de datos se logra diseñando una estrategia adecuada para manejar ruido, valores incompletos, secuencias de tiempo, casos extremos (si es necesario), etc.
- Seleccionar y aplicar el método de minería de datos apropiado: esto incluye la selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc. La selección de él o de los algoritmos a utilizar. La transformación de los datos al formato requerido por el algoritmo específico de minería de datos. Y llevar a cabo el proceso de minería de datos, se buscan patrones que puedan

expresarse como un modelo o simplemente que expresen dependencias de los datos, el modelo encontrado depende de su función (clasificación) y de su forma de representarlo (árboles de decisión, reglas, etc.), se tiene que especificar un criterio de preferencia para seleccionar un modelo dentro de un conjunto posible de modelos, se tiene que especificar la estrategia de búsqueda a utilizar (normalmente está predeterminada en el algoritmo de minería)

- Evaluación, interpretación, transformación y representación de los patrones extraídos:

Interpretar los resultados y posiblemente regresar a los pasos anteriores. Esto puede involucrar repetir el proceso, quizás con otros datos, otros algoritmos, otras metas y otras estrategias. Este es un paso crucial en donde se requiere tener conocimiento del dominio. La interpretación puede beneficiarse de procesos de visualización, y sirve también para borrar patrones redundantes oirrelevantes.

- Difusión y uso del nuevo conocimiento.

Incorporar el conocimiento descubierto al sistema (normalmente para mejorarlo) lo cual puede incluir resolver conflictos potenciales con el conocimiento existente.

El conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente para almacenarlo y reportarlo a las personas interesadas.

En este sentido, KDD implica un proceso interactivo e iterativo involucrando la aplicación de varios algoritmos de minería de datos.

Capítulo 3

Mineria de Datos - Data Mining

3.1 Conceptos e Historia

Aunque desde un punto de vista académico el término data mining es una etapa dentro de un proceso mayor llamado extracción de conocimiento en bases de datos, (mencionado en el capitulo anterior) en el entorno comercial, así como en este trabajo, ambos términos se usan de manera indistinta. Lo que en verdad hace el data mining es reunir las ventajas de varias áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo, principalmente usando como materia prima las bases de datos. Una definición tradicional es la siguiente: *Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos* (Fayyad y otros, 1996). Desde el punto de vista empresarial , lo definimos como: *La integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión* (Molina y otros, 2001).

La idea de data mining no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como data fishing, data mining o data archaeology con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a

consolidar los términos de data mining y KDD.[3] A finales de los años ochenta sólo existían un par de empresas dedicadas a esta tecnología; en 2002 existen más de 100 empresas en el mundo que ofrecen alrededor de 300 soluciones. Las listas de discusión sobre este tema las forman investigadores de más de ochenta países. Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios.

El data mining es una tecnología compuesta por etapas que integra varias áreas y que no se debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones software en cada etapa que pueden ser estadísticas, de visualización de datos o de inteligencia artificial, principalmente. Actualmente existen aplicaciones o herramientas comerciales de data mining muy poderosas que contienen un sinfín de utilerías que facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta.

La data mining es la etapa de descubrimiento en el proceso de KDD: *Paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos preprocesados* (Fayyad et al., 1996)Aunque se suelen usar indistintamente los términos KDD y Minería de Datos.

3.1.1 Los Fundamentos del Data Mining

Las técnicas de Data Mining son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos de negocios fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. Data Mining toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, hacia la entrega de información prospectiva y proactiva. Data Mining está listo para su aplicación en la comunidad de negocios porque está soportado por tres tecnologías que ya están suficientemente maduras:

- Recolección masiva de datos.
- Potentes computadoras con multiprocesadores.
- Algoritmos de Data Mining.

Las bases de datos comerciales están creciendo a un ritmo sin precedentes. Un reciente estudio del META GROUP sobre los proyectos de Data Warehouse encontró que el 19% de los que contestaron están por encima del nivel de los 50 Gigabytes, mientras que el 59% espera alcanzarlo en el segundo trimestre de 1997. En algunas industrias, tales como ventas al por menor (retail), estos números pueden ser aún mayores. MCI Telecommunications Corp. cuenta con una base de datos de 3 terabytes + 1 terabyte de índices y overhead corriendo en MVS sobre IBM SP2. La necesidad paralela de motores computacionales mejorados puede ahora alcanzarse de forma más costo - efectiva con tecnología de computadoras con multiprocesamiento paralelo. Los algoritmos de Data Mining utilizan técnicas que han existido por lo menos desde hace 10 años, pero que sólo han sido implementadas recientemente como herramientas maduras, confiables, entendibles que consistentemente son más performantes que métodos estadísticos clásicos.

En la evolución desde los datos de negocios a información de negocios, cada nuevo paso se basa en el previo. Por ejemplo, el acceso a datos dinámicos es crítico para las aplicaciones de navegación de datos (drill through applications), y la habilidad para almacenar grandes bases de datos es crítica para Data Mining.

Los componentes esenciales de la tecnología de Data Mining han estado bajo desarrollo por décadas, en áreas de investigación como estadísticas, inteligencia artificial y aprendizaje de máquinas. Hoy, la madurez de estas técnicas, junto con los motores de bases de datos relacionales de alta performance, hicieron que estas tecnologías fueran prácticas para los entornos de data warehouse actuales.

3.2 Principales características y objetivos de la Minería de Datos

- Explorar los datos se encuentran en las profundidades de las bases de datos, como los almacenes de datos, que algunas veces contienen información almacenada durante varios años.

- En algunos casos, los datos se consolidan en un almacén de datos y en mercados de datos; en otros, se mantienen en servidores de Internet e Intranet.

- El entorno de la minería de datos suele tener una arquitectura cliente-servidor.
- Las herramientas de la minería de datos ayudan a extraer el mineral de la información enterrado en archivos corporativos o en registros públicos, archivados
- El minero es, muchas veces un usuario final con poca o ninguna habilidad de programación, facultado por barrenadoras de datos y otras poderosas herramientas indagatorias para efectuar preguntas adhoc y obtener rápidamente respuestas.
- Hurgar y sacudir a menudo implica el descubrimiento de resultados valiosos e inesperados.
- Las herramientas de la minería de datos se combinan fácilmente y pueden analizarse y procesarse rápidamente.
- Debido a la gran cantidad de datos, algunas veces resulta necesario usar procesamiento en paralelo para la minería de datos.
- La minería de datos produce cinco tipos de información:
  - Asociaciones.
  - Secuencias.
  - Clasificaciones.
  - Agrupamientos.
  - Pronósticos.
- Los mineros de datos usan varias herramientas y técnicas.

La minería de datos es un proceso que invierte la dinámica del método científico en el siguiente sentido:

En el método científico, primero se formula la hipótesis y luego se diseña el experimento para coleccionar los datos que confirmen o refuten la hipótesis. Si esto se hace con la formalidad adecuada (cuidando cuáles son las variables controladas y cuáles experimentales), se obtiene un nuevo conocimiento.

En la minería de datos, se coleccionan los datos y se espera que de ellos emerjan hipótesis. Se busca que los datos describan o indiquen por qué son

como son. Luego entonces, se valida esa hipótesis inspirada por los datos en los datos mismos, será numéricamente significativa, pero experimentalmente inválida. De ahí que la minería de datos debe presentar un enfoque exploratorio, y no confirmador. Usar la minería de datos para confirmar las hipótesis formuladas puede ser peligroso, pues se está haciendo una inferencia poco válida.

La minería de datos es una tecnología compuesta por etapas que integra varias áreas y que no se debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones software en cada etapa que pueden ser estadísticas, de visualización de datos o de inteligencia artificial, principalmente. Actualmente existen aplicaciones o herramientas comerciales de minería de datos muy poderosas que contienen un sinfín de utilerías que facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta.

3.3 El Alcance de Data Mining

El nombre de Data Mining deriva de las similitudes entre buscar valiosa información de negocios en grandes bases de datos - por ej.: encontrar información de la venta de un producto entre grandes montos de Gigabytes almacenados - y minar una montaña para encontrar una veta de metales valiosos. Ambos procesos requieren examinar una inmensa cantidad de material, o investigar inteligentemente hasta encontrar exactamente donde residen los valores. Dadas bases de datos de suficiente tamaño y calidad, la tecnología de Data Mining puede generar nuevas oportunidades de negocios al proveer estas capacidades:

- Predicción automatizada de tendencias y comportamientos. Data Mining automatiza el proceso de encontrar información predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos. Un típico ejemplo de problema predecible es el marketing apuntado a objetivos (targeted marketing). Data Mining usa datos en mailing promocionales anteriores para identificar posibles objetivos para maximizar los resultados de la inversión en futuros mailing. Otros problemas predecibles incluyen pronósticos de problemas financieros futuros y otras formas de incumplimiento, e identificar segmentos de población que probablemente respondan similarmente a eventos dados.

- Descubrimiento automatizado de modelos previamente desconocidos. Las herramientas de Data Mining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso. Otros problemas de descubrimiento de modelos incluye detectar transacciones fraudulentas de tarjetas de créditos e identificar datos anormales que pueden representar errores de tipeado en la carga de datos.

Las técnicas de Data Mining pueden redituar los beneficios de automatización en las plataformas de hardware y software existentes y puede ser implementadas en sistemas nuevos a medida que las plataformas existentes se actualicen y nuevos productos sean desarrollados. Cuando las herramientas de Data Mining son implementadas en sistemas de procesamiento paralelo de alta performance, pueden analizar bases de datos masivas en minutos. Procesamiento más rápido significa que los usuarios pueden automáticamente experimentar con más modelos para entender datos complejos. Alta velocidad hace que sea práctico para los usuarios analizar inmensas cantidades de datos. Grandes bases de datos, a su vez, producen mejores predicciones.

3.4 Una arquitectura para Data Mining

Para aplicar mejor estas técnicas avanzadas, éstas deben estar totalmente integradas con el data warehouse así como con herramientas flexibles e interactivas para el análisis de negocios. Varias herramientas de Data Mining actualmente operan fuera del warehouse, requiriendo pasos extra para extraer, importar y analizar los datos. Además, cuando nuevos conceptos requieren implementación operacional, la integración con el warehouse simplifica la aplicación de los resultados desde Data Mining. El Data warehouse analítico resultante puede ser aplicado para mejorar procesos de negocios en toda la organización, en áreas tales como manejo de campañas promocionales, detección de fraudes, lanzamiento de nuevos productos, etc.

El punto de inicio ideal es un data warehouse que contenga una combinación de datos de seguimiento interno de todos los clientes junto con datos externos de mercado acerca de la actividad de los competidores. Información histórica sobre potenciales clientes también provee una excelente base para prospecting. Este warehouse puede ser implementado en una variedad de sistemas de bases relacionales y debe ser optimizado para un acceso a los datos flexible y rápido.

Capítulo 4

Fases de un Proyecto de MD y Aplicaciones de Uso

4.1 Fases de un Proyecto de Minería de Datos

Los pasos a seguir para la realización de un proyecto de minería de datos son siempre los mismos, independientemente de la técnica específica de extracción de conocimiento usada. En la sigiente figura 4.1 de la pagina 18 se ilustra las Fases del Proyecto de MD.

El proceso de minería de datos pasa por las siguientes fases:

- Filtrado de datos.
- Selección de Variables.
- Extracción de Conocimiento.
- Interpretación y Evaluación.

4.1.1 Filtrado de datos

El formato de los datos contenidos en la fuente de datos (base de datos, Data Warehouse...) nunca es el idóneo, y la mayoría de las veces no es posible ni siquiera utilizar ningún algoritmo de minería sobre los datos en bruto.

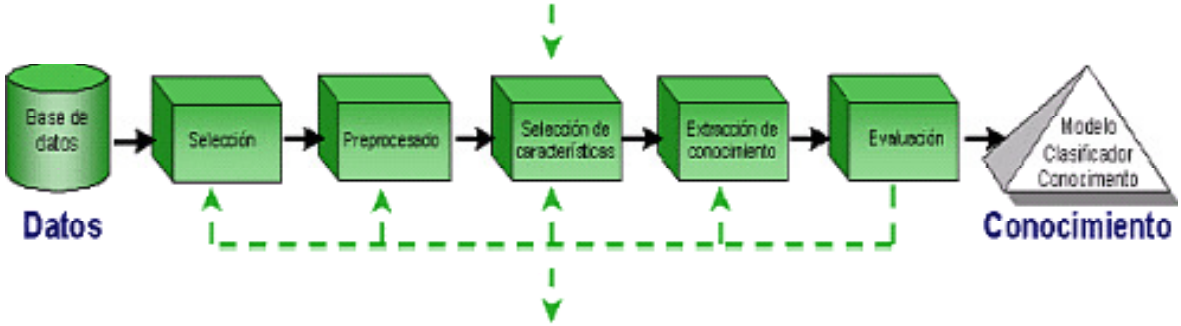


Figura 4.1: Fases del Proyecto de M.D.

Mediante el preprocesado, se filtran los datos (de forma que se eliminan valores incorrectos, no válidos, desconocidos... según las necesidades y el algoritmo a usar), se obtienen muestras de los mismos (en busca de una mayor velocidad de respuesta del proceso), o se reducen el número de valores posibles (mediante redondeo, clustering,...).

4.1.2 Selección de variables

Aún después de haber sido preprocesados, en la mayoría de los casos se tiene una cantidad ingente de datos. La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería.

Los métodos para la selección de características son básicamente dos:

- Aquellos basados en la elección de los mejores atributos del problema,
- Y aquellos que buscan variables independientes mediante tests de sensibilidad, algoritmos de distancia o heurísticos.

4.1.3 Algoritmos de Extracción de Conocimiento

Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores

de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.

4.1.4 Interpretación y evaluación

Una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

4.2 Aplicaciones de Uso

Cada año, en los diferentes congresos, simposios y talleres que se realizan en el mundo se reúnen investigadores con aplicaciones muy diversas. Sobre todo en los Estados Unidos, el data mining se ha ido incorporando a la vida de empresas, gobiernos, universidades, hospitales y diversas organizaciones que están interesadas en explorar sus bases de datos.

4.2.1 En el Gobierno:

- **El FBI analizará las bases de datos comerciales para detectar terroristas.**

A principios del mes de julio de 2002, el director del Federal Bureau of Investigation (FBI), John Aschcroft, anunció que el Departamento de Justicia comenzará a introducirse en la vasta cantidad de datos comerciales referentes a los hábitos y preferencias de compra de los consumidores, con el fin de descubrir potenciales terroristas antes de que ejecuten una acción. Algunos expertos aseguran que, con esta información, el FBI unirá todas las bases de datos probablemente mediante el número de la Seguridad Social y permitirá saber si una persona fuma, qué talla y tipo de ropa usa, su registro de arrestos,

su salario, las revistas a las que está suscrito, su altura y peso, sus contribuciones a la Iglesia, grupos políticos u organizaciones no gubernamentales, sus enfermedades crónicas (como diabetes o asma), los libros que lee, los productos de supermercado que compra, si tomó clases de vuelo o si tiene cuentas de banco abiertas, entre otros. La inversión inicial ronda los setenta millones de dólares estadounidenses para consolidar los almacenes de datos, desarrollar redes de seguridad para compartir información e implementar nuevo software analítico y de visualización.

4.2.2 En la Empresa

- **Detección de fraudes en las tarjetas de crédito.**

  En 2001, las instituciones financieras a escala mundial perdieron más de 2.000 millones de dólares estadounidenses en fraudes con tarjetas de crédito y débito. El Falcon Fraud Manager es un sistema inteligente que examina transacciones, propietarios de tarjetas y datos financieros para detectar y mitigar fraudes. En un principio estaba pensado, en instituciones financieras de Norteamérica, para detectar fraudes en tarjetas de crédito. Sin embargo, actualmente se le han incorporado funcionalidades de análisis en las tarjetas comerciales, de combustibles y de débito. El sistema Falcon ha permitido ahorrar más de seiscientos millones de dólares estadounidenses cada año y protege aproximadamente más de cuatrocientos cincuenta millones de pagos con tarjeta en todo el mundo -aproximadamente el sesenta y cinco por ciento de todas las transacciones con tarjeta de crédito.

- **Descubriendo el porqué de la deserción de clientes de una compañía operadora de telefonía móvil.**

  Este estudio fue desarrollado en una operadora española que básicamente situó sus objetivos en dos puntos: el análisis del perfil de los clientes que se dan de baja y la predicción del comportamiento de sus nuevos clientes. Se analizaron los diferentes históricos de clientes que habían abandonado la operadora (12,6%) y de clientes que continuaban con su servicio (87,4%). También se analizaron las variables personales de cada cliente (estado civil, edad, sexo, nacionalidad, etc.). De igual forma se estudiaron, para cada cliente, la morosidad, la frecuencia y el horario de uso del servicio, los descuentos y el porcentaje de llamadas locales, interprovinciales, internacionales y gratuitas. Al contrario de lo que se podría pensar, los clientes que abandonaban la operadora generaban ganancias

para la empresa; sin embargo, una de las conclusiones más importantes radicó en el hecho de que los clientes que se daban de baja recibían pocas promociones y registraban un mayor número de incidencias respecto a la media. De esta forma se recomendó a la operadora hacer un estudio sobre sus ofertas y analizar profundamente las incidencias recibidas por esos clientes. Al descubrir el perfil que presentaban, la operadora tuvo que diseñar un trato más personalizado para sus clientes actuales con esas características. Para poder predecir el comportamiento de sus nuevos clientes se diseñó un sistema de predicción basado en la cantidad de datos que se podía obtener de los nuevos clientes comparados con el comportamiento de clientes anteriores.

- **Hábitos de compra en supermercados.**

  Un estudio muy citado detectó que los viernes había una cantidad inusualmente elevada de clientes que adquirían a la vez pañales y cerveza. Se detectó que se debía a que dicho día solían acudir al supermercado padres jóvenes cuya perspectiva para el fin de semana consistía en quedarse en casa cuidando de su hijo y viendo la televisión con una cerveza en la mano. El supermercado pudo incrementar sus ventas de cerveza colocándolas próximas a los pañales para fomentar las ventas compulsivas,

- **Prediciendo el tamaño de las audiencias televisivas.**

  La British Broadcasting Corporation (BBC) del Reino Unido emplea un sistema para predecir el tamaño de las audiencias televisivas para un programa propuesto, así como el tiempo óptimo de exhibición (Brachman y otros, 1996). El sistema utiliza redes neuronales y árboles de decisión aplicados a datos históricos de la cadena para determinar los criterios que participan según el programa que hay que presentar. La versión final se desempeña tan bien como un experto humano con la ventaja de que se adapta más fácilmente a los cambios porque es constantemente reentrenada con datos actuales.

4.2.3 En la Universidad

- **Conociendo si los recién titulados de una universidad llevan a cabo actividades profesionales relacionadas con sus estudios.**

Se hizo un estudio sobre los recién titulados de la carrera de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Chihuahua II, en Méjico (Rodas, 2001). Se quería observar si sus recién titulados se insertaban en actividades profesionales relacionadas con sus estudios y, en caso negativo, se buscaba saber el perfil que caracterizó a los exalumnos durante su estancia en la universidad. El objetivo era saber si con los planes de estudio de la universidad y el aprovechamiento del alumno se hacía una buena inserción laboral o si existían otras variables que participaban en el proceso. Dentro de la información considerada estaba el sexo, la edad, la escuela de procedencia, el desempeño académico, la zona económica donde tenía su vivienda y la actividad profesional, entre otras variables. Mediante la aplicación de conjuntos aproximados se descubrió que existían cuatro variables que determinaban la adecuada inserción laboral, que son citadas de acuerdo con su importancia: zona económica donde habitaba el estudiante, colegio de donde provenía, nota al ingresar y promedio final al salir de la carrera. A partir de estos resultados, la universidad tendrá que hacer un estudio socioeconómico sobre grupos de alumnos que pertenecían a las clases económicas bajas para dar posibles soluciones, debido a que tres de las cuatro variables no dependían de la universidad.

4.2.4 En Investigaciones Espaciales

- **Proyecto SKYCAT.**

  Durante seis años, el Second Palomar Observatory Sky Survey (POSS-II) coleccionó tres terabytes de imágenes que contenían aproximadamente dos millones de objetos en el cielo. Tres mil fotografías fueron digitalizadas a una resolución de 16 bits por píxel con 23.040 x 23.040 píxeles por imagen. El objetivo era formar un catálogo de todos esos objetos. El sistema Sky Image Cataloguing and Analysis Tool (SKYCAT) se basa en técnicas de agrupación (clustering) y árboles de decisión para poder clasificar los objetos en estrellas, planetas, sistemas, galaxias, etc. con una alta confiabilidad (Fayyad y otros, 1996). Los resultados han ayudado a los astrónomos a descubrir dieciséis nuevos quásars con corrimiento hacia el rojo que los incluye entre los objetos más lejanos del universo y, por consiguiente, más antiguos. Estos quásars son difíciles de encontrar y permiten saber más acerca de los orígenes del universo.

4.2.5 En los Clubes Deportivos

- **Los equipos de la NBA utilizan aplicaciones inteligentes para apoyar a su cuerpo de entrenadores.**

  El Advanced Scout es un software que emplea técnicas de data mining y que han desarrollado investigadores de IBM para detectar patrones estadísticos y eventos raros. Tiene una interfaz gráfica muy amigable orientada a un objetivo muy específico: analizar el juego de los equipos de la National Basketball Association (NBA).

  El software utiliza todos los registros guardados de cada evento en cada juego: pases, encestes, rebotes y doble marcaje (double team) a un jugador por el equipo contrario, entre otros. El objetivo es ayudar a los entrenadores a aislar eventos que no detectan cuando observan el juego en vivo o en película.

  Un resultado interesante fue uno hasta entonces no observado por los entrenadores de los Knicks de Nueva York. El doble marcaje a un jugador puede generalmente dar la oportunidad a otro jugador de encestar más fácilmente. Sin embargo, cuando los Bulls de Chicago jugaban contra los Knicks, se encontró que el porcentaje de encestes después de que al centro de los Knicks, Patrick Ewing, le hicieran doble marcaje era extremadamente bajo, indicando que los Knicks no reaccionaban correctamente a los dobles marcajes. Para saber el porqué, el cuerpo de entrenadores estudió cuidadosamente todas las películas de juegos contra Chicago. Observaron que los jugadores de Chicago rompían su doble marcaje muy rápido de tal forma que podían tapar al encestador libre de los Knicks antes de prepararse para efectuar su tiro. Con este conocimiento, los entrenadores crearon estrategias alternativas para tratar con el doble marcaje.

  La temporada pasada, IBM ofreció el Advanced Scout a la NBA, que se convirtió así en un patrocinador corporativo. La NBA dio a sus veintinueve equipos la oportunidad de aplicarlo. Dieciocho equipos lo están haciendo hasta el momento obteniendo descubrimientos interesantes.

Capítulo 5

Extensiones del Data Mining

5.1 Web mining

Una de las extensiones del data mining consiste en aplicar sus técnicas a documentos y servicios del Web, lo que se llama web mining (minería de web) (Kosala y otros, 2000). Todos los que visitan un sitio en Internet dejan huellas digitales (direcciones de IP, navegador, galletas, etc.) que los servidores automáticamente almacenan en una bitácora de accesos (log). Las herramientas de web mining analizan y procesan estos logs para producir información significativa, por ejemplo, cómo es la navegación de un cliente antes de hacer una compra en línea. Debido a que los contenidos de Internet consisten en varios tipos de datos, como texto, imagen, vídeo, metadatos o hiperligas, investigaciones recientes usan el término multimedia data mining (minería de datos multimedia) como una instancia del web mining (Zaiane y otros, 1998) para tratar ese tipo de datos. Los accesos totales por dominio, horarios de accesos más frecuentes y visitas por día, entre otros datos, son registrados por herramientas estadísticas que complementan todo el proceso de análisis del web mining.

Normalmente, el web mining puede clasificarse en tres dominios de extracción de conocimiento de acuerdo con la naturaleza de los datos:

- Web content mining (minería de contenido web). Es el proceso que consiste en la extracción de conocimiento del contenido de documentos o sus descripciones. La localización de patrones en el texto de los documen-

tos, el descubrimiento del recurso basado en conceptos de indexación o la tecnología basada en agentes también pueden formar parte de esta categoría.

- Web structure mining (minería de estructura web). Es el proceso de inferir conocimiento de la organización del WWW y la estructura de sus ligas.
- Web usage mining (minería de uso web). Es el proceso de extracción de modelos interesantes usando los logs de los accesos al web.

Algunos de los resultados que pueden obtenerse tras la aplicación de los diferentes métodos de web mining son:

- El ochenta y cinco por ciento de los clientes que acceden a la página home de productos y a la de noticias de la misma página acceden también a la página de historia. Esto podría indicar que existe alguna noticia interesante de la empresa que hace que los clientes se dirijan a historias de suceso. Igualmente, este resultado permitiría detectar la noticia sobresaliente y colocarla quizá en la página principal de la empresa.
- El sesenta por ciento de los clientes que hicieron una compra en línea en la página del producto 1 también compraron en la página del producto 4 después de un mes. Esto indica que se podría recomendar en la página del producto 1 comprar el producto 4 y ahorrarse el costo de envío de este producto.

Los anteriores ejemplos ayudan a formar una pequeña idea de lo que se puede obtener. Sin embargo, en la realidad existen herramientas de mercado muy poderosas con métodos variados y visualizaciones gráficas excelentes.

5.2 Text mining

Estudios recientes indican que el ochenta por ciento de la información de una compañía está almacenada en forma de documentos. Sin duda, este campo de estudio es muy vasto, por lo que técnicas como la categorización de texto, el procesamiento de lenguaje natural, la extracción y recuperación de la información o el aprendizaje automático, entre otras, apoyan al text mining (minería

de texto). En ocasiones se confunde el text mining con la recuperación de la información (Information Retrieval o IR) (Hearst, 1999). Ésta última consiste en la recuperación automática de documentos relevantes mediante indexaciones de textos, clasificación, categorización, etc. Generalmente se utilizan palabras clave para encontrar una página relevante. En cambio, el text mining se refiere a examinar una colección de documentos y descubrir información no contenida en ningún documento individual de la colección; en otras palabras, trata de obtener información sin haber partido de algo (Nasukawa y otros, 2001).

Una aplicación muy popular del text mining es relatada en Hearst (1999). Don Swanson intenta extraer información derivada de colecciones de texto. Teniendo en cuenta que los expertos sólo pueden leer una pequeña parte de lo que se publica en su campo, por lo general no se dan cuenta de los nuevos desarrollos que se suceden en otros campos. Así, Swanson ha demostrado cómo cadenas de implicaciones causales dentro de la literatura médica pueden conducir a hipótesis para enfermedades poco frecuentes, algunas de las cuales han recibido pruebas de soporte experimental. Investigando las causas de la migraña, dicho investigador extrajo varias piezas de evidencia a partir de títulos de artículos presentes en la literatura biomédica. Algunas de esas claves fueron:

- El estrés está asociado con la migraña.
- El estrés puede conducir a la pérdida de magnesio.
- Los bloqueadores de canales de calcio previenen algunas migrañas.
- El magnesio es un bloqueador natural del canal de calcio.
- La depresión cortical diseminada (DCD) está implicada en algunas migrañas.
- Los niveles altos de magnesio inhiben la DCD.
- Los pacientes con migraña tienen una alta agregación plaquetaria.
- El magnesio puede suprimir la agregación plaquetaria.

Estas claves sugieren que la deficiencia de magnesio podría representar un papel en algunos tipos de migraña, una hipótesis que no existía en la literatura y que Swanson encontró mediante esas ligas. De acuerdo con Swanson (Swanson y otros, 1994), estudios posteriores han probado experimentalmente esta hipótesis obtenida por text mining con buenos resultados.

Capítulo 6

Conclusión

Nuestra capacidad para almacenar datos ha crecido en los últimos años a velocidades exponenciales. En contrapartida, nuestra capacidad para procesarlos y utilizarlos no ha ido a la par. Por este motivo, el data mining se presenta como una tecnología de apoyo para explorar, analizar, comprender y aplicar el conocimiento obtenido usando grandes volúmenes de datos. Descubrir nuevos caminos que nos ayuden en la identificación de interesantes estructuras en los datos es una de las tareas fundamentales en el data mining.

En el ámbito comercial, resulta interesante encontrar patrones ocultos de consumo de los clientes para poder explorar nuevos horizontes. Saber que un vehículo deportivo corre un riesgo de accidente casi igual al de un vehículo normal cuando su dueño tiene un segundo vehículo en casa ayuda a crear nuevas estrategias comerciales para ese grupo de clientes. Asimismo, predecir el comportamiento de un futuro cliente, basándose en los datos históricos de clientes que presentaron el mismo perfil, ayuda a poder retenerlo durante el mayor tiempo posible.

Las herramientas comerciales de data mining que existen actualmente en el mercado son variadas y excelentes. Las hay orientadas al estudio del web o al análisis de documentos o de clientes de supermercado, mientras que otras son de uso más general. Su correcta elección depende de la necesidad de la empresa y de los objetivos a corto y largo plazo que pretenda alcanzar. La decisión de seleccionar una solución de data mining no es una tarea simple. Es necesario consultar a expertos en el área con vista a seleccionar la más adecuada para el problema de la empresa.

Como se ha visto a lo largo del este artículo, son muchas las áreas, técnicas, estrategias, tipos de bases de datos y personas que intervienen en un proceso de data mining. Los negocios requieren que las soluciones tengan una integración transparente en un ambiente operativo. Esto nos lleva a la necesidad de establecer estándares para hacer un ambiente interoperable, eficiente y efectivo. Se exponen algunas iniciativas para estos estándares, incluyendo aspectos en:

- Modelos: para representar datos estadísticos y de data mining.
- Atributos: para representar la limpieza, transformación y agregación de atributos usados como entrada en los modelos.
- Interfaces y API: para facilitar la integración con otros lenguajes o aplicaciones de software y API.
- Configuración: para representar parámetros internos requeridos para construir y usar los modelos.
- Procesos: para producir, desplegar y usar modelos.
- Datos remotos y distribuidos: para analizar y explorar datos remotos y distribuidos.

En resumen, el data mining se presenta como una tecnología emergente, con varias ventajas: por un lado, resulta un buen punto de encuentro entre los investigadores y las personas de negocios; por otro, ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios. Además, no hay duda de que trabajar con esta tecnología implica cuidar un sinnúmero de detalles debido a que el producto final involucra toma de decisiones.

Bibliografía

[1] Jhon Wiley Alan Simon and Sons. *Data Warehouse, Data Mining and OLAP*. USA, 1997.

[2] Mc Graw Hill Alex Berson, Stephen J. Smith. *Data Warehouse, Data Mining and OLAP*. USA, 1997.

[3] María José Ramírez Quintana José Hernández Orallo. *Extracción Automática de Conocimiento en Bases de Datos e Ingeniería del Software*. España, 2003.

[4] IBM Press. *IBM DB2 Intelligent Miner for Data: Utilización del Visualizador de Asociaciones*. IBM Press, USA, 1999.

[5] IBM Press. *IBM DB2 Warehouse Manager Guía de Instalación Version 7*. IBM Press, USA, 2001.

[6] IBM Press. *IBM DB2 Intelligent Miner for Data: Utilización de Intelligent Miner for Data*. IBM Press, USA, 2002.

[7] IBM Press. *IBM DB2 Intelligent Miner Visualization: Using the Intelligent Miner Visualizers*. IBM Press, USA, 2002.

[8] Colin J. White. *IBM Enterprise Analytics for the Intelligent e-Business*. IBM Press, USA, 2001.