Universidad Autónoma del Estado de México

CENTRO UNIVERSITARIO UAEM TEXCOCO

"MINERÍA DE DATOS: MÉTODO PARA OBTENER Y EXPLORAR INFORMACIÓN EN GRANDES CANTIDADES DE DATOS"

T E S I N A

PARA OBTENER EL TITULO DE

LICENCIADO EN

INFORMATICA ADMINISTRATIVA

PRESENTA:

CARLOS DÍAZ SÁNCHEZ

DIRECTOR:

DR. EN C. ADRIÁN TRUEBA ESPINOSA

REVISORES:

L. EN I.A. FABIOLA MARTÍNEZ MEJÍA

M. EN C. C. ÁNGEL RAFAEL QUINTOS RAMÍREZ

TEXCOCO, ESTADO DE MEXICO NOVIEMBRE 2012

AGRADECIMIENTOS

A mi madre, Rosa María Sánchez Estrada que siempre ha dado su apoyo incondicional, a quien le debó la vida y este triunfo profesional, por todo su trabajo y dedicación para darme una formación académica y sobre todo humanística y espiritual. De ella es este triunfo y todo mi agradecimiento.

Así mismo, agradezco a mi director de tesis Dr. Adrian Trueba Espinosa, por la paciencia y dedicarme parte de su valioso tiempo en la elaboración de este trabajo, también por compartir sus conocimientos que me sirvieron de gran ayuda.

A mis asesores, L. en I. Fabiola Martínez Mejía, M. en C.C. Ángel Rafael Quintos Ramírez por su colaboración y su gran apoyo que me dieron en este trabajo de investigación. Les agradezco bastante.

A todos mis familiares y a todas aquellas personas que han sido importantes para mí durante todo este tiempo. A todos mis maestros que aportaron a mi formación. Para quienes me enseñaron más que el saber científico. A quienes me enseñaron a ser lo que no se aprende en el salón de clase y a compartir el conocimiento con los demás.

¡GRACIAS!

Carlos Díaz Sánchez

CONTENIDO GENERAL

CONTENIDO DE FIGURAS

CONTENIDO DE TABLAS

I.-INTRODUCCION

El uso masivo de las bases de datos y los avances para aumentar las capacidades de almacenamiento de datos, han hecho que todo tipo de organizaciones puedan disponer de grandes cantidad de datos relativos a la actividad de una empresa. En muchas de estas organizaciones se han dado cuenta del potencial que tienen estos datos para conocer el comportamiento y movimiento de los productos que promueve o del personal que manejan y/o de muchas cosas más.

El análisis de los datos permite ver la evolución y desarrollo de las organizaciones, y por lo tanto, trazar un plan que permita proyectar por donde moverse en un futuro. Así, el estudio de los datos para obtener información ofrece una visión de qué se está haciendo y cómo se están haciendo los procesos y prospectiva (cómo puede evolucionar la organización en un futuro a corto-medio plazo) de la organización, y es por ello por lo que tiene una función de apoyo a la toma de decisiones.
El análisis de las bases de datos ha sido, y es, común en organizaciones económicas y empresariales, desde supermercados hasta grandes multinacionales, pero también en organismos científicos que manejan grandes cantidades de información se ha visto la utilidad de este tipo de estudios.

Para este estudio o análisis de los datos se han desarrollado técnicas que se agrupan en la minería de datos. Considerando la importancia actual de estos métodos de análisis en este trabajo se hará una investigación documental para conocer sobre el proceso de manejo de datos y de sus beneficios.

II.-PLANTEAMIENTO DEL PROBLEMA

La minería de datos es una tecnología que en los últimos años está siendo utilizada para la toma de decisiones en muchos organismos públicos y privados, sin embargo, estas técnicas no han sido masificadas en las instituciones educativas a nivel licenciatura, lo que hace que de alguna manera los estudiantes queden en desventaja con otros, que en términos generales seria un problema.

II.-JUSTIFICACION

La minería de datos se aplica a todo los datos que se puedan obtener: desde datos numéricos a imágenes de satélite, mamografías, música, etc. Se puede decir que "cualquier cosa" constituye un dato. Por tanto la minería de datos tiene infinitas aplicaciones: comerciales, marketing, industria, internet, agricultura entre otras. Conocerla es importante, por tal motivo en este trabajo se hará una investigación documental que permita a los estudiantes de licenciatura de esta institución, obtener un concentrado de documentos que le permitan conocer sobre esta disciplina que se está utilizando en muchas aplicaciones.

IV. OBJETIVOS

4.1 Objetivo general

- Realizar una investigación documental sobre la importancia y aplicaciones que tienen la minería de datos

4.2 Objetivos específicos

- Documentar las herramientas y técnicas de algoritmos sofisticados que se aplican sobre un conjunto de cantidades para obtener resultados de la extracción de información a partir de grandes bases de datos.
- Determinar las ventajas y desventajas de las diferentes técnicas que se utilizan en la minería de datos.

  Identificar las principales aplicaciones de la minería de datos.

V. MARCO TEORICO

5.1 ORÍGENES DE LA MINERÍA DE DATOS

El data mining o la minería de datos como le llamaremos desde ahora, surge como una tecnología que intenta ayudar comprender el contenido de una base de datos.

Este concepto apareció hace más de 10 años. El interés en este campo y su explotación en diferentes especialidades (negocios, finanzas, ingeniería, banca, salud, sistemas de energía, meteorología), se ha incrementado debido a la combinación de datos.
Surgimiento de la gran cantidad de datos (terabyrtes-1012 bytes – de datos) debido a la medición y recopilación de datos automática registros digitales. Archivos centralizados de datos y simulación de software y hardware (DIÍAZ, 2002).

Desde los años sesenta los estadísticos manejaban términos como este (minería de datos) con la idea de encontrar una posible solución previa en una base de datos con ruido.
A finales de los años 80´s 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 (MOLINA, 2002a).

Hasta hace poco años era una técnica experimental, con el desarrollo de Internet se ha potencializado su uso y actualmente es una técnica utilizada en el comercio electrónico. La novedad no radica en la técnica del cálculo, si no en la aplicación interactiva de la misma.

Existen tres razones fundamentales por las cuales la minería de datos es una realidad en nuestros días:

- Avances tecnológicos en almacenamiento masivo de datos y C.P.U.
- Existencia de nuevos algoritmos para extraer información en forma eficiente.
- Existencia de herramientas automáticas que no hacen necesario el ser un experto en estadística, redes neuronales, y algoritmos matemáticos para convertirse en un minero de datos (CABENA.1997).

5.2 FUNDAMENTOS

5.2.1 Definiciones

La minería de datos está siendo utilizada cada día más en los negocios, por lo tanto, se ha vuelto un área de oportunidad para nuevas generaciones.

"El descubrimiento eficiente de información valiosa, no-obvia de una gran colección de "datos" (BIGUS, 1996).

Desde un punto de vista estrecho se define como: "El descubrimiento automático de patrones interesantes y no obvios escondidos en una base de datos, los cuales tiene un gran potencial para contribuir en los aspectos principales del negocio" (PEACOCK, 1998).

"La minería de datos es un proceso no trivial de identificación válida, novedosa potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos" (FAYYAD, 1996).

"Es la integración de un conjunto de áreas que tiene como compromiso la identificación de un conocimiento obtenido a partir de la base de datos que aporten un riesgo hacia la toma de decisión" (MOLINA, 2001b).

"La minería de datos, es una técnica que permite la extracción de conocimiento útil de grandes bases de datos, previamente desconocidos, mediante el análisis de la información".

La minería de datos puede ser dividida en:

- Minería de datos predictiva: usa técnicas estadísticas.
- Minería de datos para descubrimiento de conocimiento. Usa técnicas de inteligencia artificial (WINTEN, 2000).

5.2.2 Funciones de la minería de datos

Es claro que antes de aplicar una técnica se deben tener bien definidas las funciones que está maneja.

- **Procesamiento de datos**: dependiendo de los objetivos y requerimientos se deben seleccionar, filtrar, agregar, extraer muestras, validar y transformar datos.
- **Predicción:** dado un conjunto de datos y un modelo de predicción que trabaja sobre ellos, se trata de predecir el valor de un atributo específico que todavía no se tiene (a veces la funcionalidad de predicción se utiliza para validar hipótesis que involucran otros datos).
- **Regresión:** es el análisis de dependencia entre valores de atributos (modelos lineales). El atributo dependiente se puede predecir aplicado el modelo de regresión y el valor de los atributos independientes.
- **Series de tiempo:** se utiliza para predecir valores de un atributo, que presenta auto correlación temporal en base a series de datos históricas del mismo atributo.
- **Clasificación:** dado un conjunto predeterminado de clases categóricas, determinar a qué clase pertenece un ítem.
- **Clustering:** divide a los datos en diferentes grupos, el objetivo es encontrar una agrupación de datos de forma que los datos de un mismo grupo sean muy similares y muy diferentes entre grupos distintos.
- **Visualización del modelo:** juegan un importante rol en KDD para la interpretación humana.
- **Análisis exploratorio de datos:** permite la exploración interactiva de datos, sin modelos preconcebidos (EDELSTEIN, 1999).

Un sistema de minería de datos debe ser capaz de:

- Describir los datos en forma resumida, dando sus principales propiedades estadísticas.
- Hacer una visualización gráfica de los datos
- Descubrir potenciales relaciones entre sus datos.

  Construir modelos predicativos, en base a los modelos encontrados.
- Verificar los modelos construidos.

Es necesario comprender las técnicas utilizadas para poder realizar un buen ajuste de parámetros y optimizar la precisión de los algoritmos utilizados (GOEBEL, 1999).

5.2.3 Técnica de la minería de datos

La técnica de la minería de datos, se emplean para mejorar el rendimiento de procesos de negocio en los que se manejan grandes bases de datos, a partir de las cuales se crean modelos predictivos y descriptivos.

- **Modelo predictivo**: responde preguntas sobre datos futuros.

¿Cuáles serán las ventas del año próximo?

¿Es esta transacción fraudulenta?

¿Qué tipo de seguro es más probable que contrate el cliente "X"?

Modelo descriptivo: proporciona información sobre las relaciones entre datos.

- Los clientes que compran pañales suelen comprar cerveza.
- El tabaco y el alcohol son los factores más importantes en la enfermedad X.
- Los clientes sin televisión y con bicicleta tienen características muy distintas del resto (HERNANDEZ, 2004).

Las técnicas más utilizadas son:

Redes neuronales. Modelos no lineales inspirados en las redes de neuronas biológicas y se usan generalmente en problemas de clasificación y predicción.

Árboles de decisión. Son estructuras en forma de árbol que representan conjuntos de decisiones capaces de generar reglas para la clasificación de los datos.

Algoritmos genéticos. Son modelos inspirados en la evolución de las especies y que se aplican generalmente en problemas de optimización. Permite incluir fácilmente ligaduras complicadas que limitan la solución a un problema.

Clustering. Métodos de agrupación de datos que nos permite clasificar los datos por su similitud entre ellos. Son utilizadas con frecuencia para entender los grupos naturales de clientes en empresas o bancos (ANSWERMATH, 2005).

	TAREAS generadoras de MODELOS				
	Predictivo (supervisado)		descriptivo (no supervisado)		
TECNICAS	Clasificación	Predicción	Cluster	Asociación	Otros
Redes Neuronales	X	X	X		
Arboles de Decisión	X	X	X		
Kohonen			X		
Regresión paramétrica		X			
Regresión Logistica	X				
K means	X		X		
Asociaciones				X	
Análisis Factorial					X
Análisis discriminante		X			

Figura 1 Técnicas en los Modelos

5.2.4 Ventajas de la minería de datos

Una empresa en posesión de una base de datos de calidad y tamaño suficiente puede emplear la minería de datos para generar nuevas oportunidades de negocio, teniendo para ella ventajas como:

Predicción automática de comportamientos

Generalmente se trata de problemas de clasificación, como ejemplo podemos citar el marketing dirigido. Minería de datos, usa los resultados de campañas de marketing realizadas anteriormente para identificar el perfil de los clientes que son más constantes al comprar el producto y de este modo permitir el correo masivo por el correo dirigido.

Predicción automática de tendencias

Se basa en una base de datos histórica, minería de datos crea un modelo para predecir las tendencias. Como ejemplos se puede citar la predicción de ventas en el futuro o la predicción en mercados de capitales.

Descubrimiento automático de comportamientos desconocidos anteriormente

Las herramientas de visualización clustering, permite ver los datos desde una perspectiva distinta y por ello descubrir nuevas relaciones entre ellos
(GARRIDO, 2001).
Según (MOLINA.2002) las ventajas son las siguientes:

- Resulta un buen punto de encuentro entre los investigadores y las personas de negocios
- Ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios
- Contribuye a la tomar de decisiones tácticas y estratégicas proporcionadas un sentido automatizado para identificar información clave desde volúmenes de datos generados por procesos tradicionales
- Permite a los usuarios dar prioridad a decisiones y acciones mostrando factores que tienden a lograr un objetivo planeado
- Permite segmentar a los clientes
- Los resultados ayudan a entender mejor el problema y entorno, siendo capaces, de definir cuál es la mejor solución

5.2.5 Métodos y metodologías de la minería de datos

Se hace una breve descripción de los métodos empleados en Minería de Datos conforme a la clasificación de Aggarwal & Yu(1998).

A.-Reglas de asociación. Este tipo de algoritmo busca encontrar correlaciones interesantes entre los datos y sus diferentes atributos, estas correlaciones son reglas que existen entre los datos y que tiene cierto grado de certeza o exactitud. Una regla de asociación como enunciado puede ser: “el 90% de los clientes que compran pan y mantequilla (antecedente) también compran leche (consecuencia)”. Su ventaja es que genera reglas de muy fácil entendimiento y su desventaja es que se requiere de muchas iteraciones.

B.-Clustering. Estos algoritmos agrupan registros similares en segmentos que tiene cierta similitud entre un grupo de puntos. Cada uno de estos segmentos debe ser tratado de manera diferente. Estos algoritmos son conocidos como algoritmos de aprendizaje supervisado y se basa ampliamente en métodos estadísticos. Su ventaja, es que pueden generar reglas claras sobre los datos sin necesidad de supervisión y su desventaja, es que al tratar de encontrar reglas que cubran óptimamente todos los clusters encontrados en un conjunto es un problema NP-hard1.

C.-Árboles de decisión. Estos clasificadores, que son supervisados, utilizan un árbol de decisión para dividir o clasificar los datos hasta que cada partición contenga la mayoría de los ejemplos de una clase. El splits point, en un árbol de decisión es un nodo en el árbol que se utiliza como condición para decidir cómo se deben dividir los datos. Las ventajas de los árboles de decisión, son que corren muy rápido y producen resultados de muy fácil interpretación. Una de sus desventajas es que pueden producirse resultados poco exactos.

D.-Vecino más cercano. Este algoritmo encuentra los vecinos más cercanos del dato de prueba y le asigna un identificador de clase de la misma que tengan sus vecinos más cercanos. También es fácil de interpretar el resultado que producen estos algoritmos, pero su desventaja es que produce una gran cantidad de reglas para un conjunto de datos pequeños.

E.-Redes Neuronales. Se basan en el modelo del cerebro humano. Es una estructura de datos de funciones a las que son dados unos pesos de entrada, producen un output que es el identificador de clase. Cada dato es alimentado en la red neuronal, las funciones van siendo modificadas conforme a los rangos de error producidos a la salida. Estos algoritmos requieren de mucho entrenamiento aun cuando los conjuntos de datos sean pequeños. La ventaja principal es que producen resultados muy exactos, y su desventaja, además los resultados no son fáciles de interpretar.

F.-Algoritmos genéticos. Estos algoritmos son utilizados para encontrar dependencias entre las variables. Las combinaciones de soluciones potenciales compiten entre ellas y la mejor solución es seleccionada y combinada con otras**.** También producen resultados exactos y su desventaja, es que los resultados que arrojan también son difíciles de comprender.

G.-Redes Bayesianas. Son grafos diseccionados a cíclicos que representan distribuciones de probabilidad. Los nodos representan atributos y estados; las aristas representan las dependencias probabilidades entre ellos. Al agregar conocimiento experto (proveniente de expertos en el dominio de aplicación) la red neuronal es modificada. Una ventaja, que tiene las redes bayesianas, es que sus resultados son muy fáciles de interpretar, la desventaja, es que las probabilidades deben ser introducidas por un experto del dominio de aplicación (AGGARWAL,1998).

Es obvio que cualquier técnica, para ser llevada a cabo necesita de una serie de pasos organizados, los cuales llevaran a la consecución de la meta establecida.

Existen dos procesos de minería de datos una que es el proceso estándar de minería de datos o CRISP-DM y el otro que es el Modelo Two Crows, ambos son similares, pero se presentan juntos para contrastar los procesos desarrollados por cada uno.

5.2.6 Proceso estándar de minería de datos

Fue definido por un grupo de compañías con amplias trayectorias en el uso de la minería de datos, este proceso consta de las siguientes fases:

Fase I

Comprensión del Problema

Los objetivos de esta fase son:

Determinación de los objetivos: el primer paso y el más importante es entender la necesidad de hacer minería de datos, determinando cual es el problema que se desea resolver, para que se convierta en el objetivo del proceso.

Los problemas pueden ser diversos: optimizar la respuesta del cliente ante una campaña publicitaria, prevenir el uso fraudulento de tarjetas de crédito, detección de intentos hostiles de entradas al sistema, etc.

Definición de criterios de éxito: una vez definido el problema, es necesario disponer de criterios de éxito. Esos criterios pueden ser objetivos (cuantitativos), por ejemplo, un mejor número de detecciones y desviaciones, una mayor respuesta de los clientes a una campaña publicitaria, un mayor porcentaje de pacientes correctamente diagnosticados. Los criterios pueden ser también subjetivos o de naturaleza cualitativa, en este caso, un experto en el área del dominio califica el resultado del esfuerzo de minería de datos con respecto al conocimiento preexistente sobre el problema. Los resultados deben contener algunas nuevas percepciones acerca de las relaciones entre las variables del dominio del problema.

Calificación de la situación: una vez definido el problema y sus criterios de la solución, hay que tomar en cuenta los aspectos relacionados al problema, como: ¿Cuál es el conocimiento experto o previo disponible acerca del problema? ¿Se tiene datos suficientes para intentar resolver el problema? ¿Se dispone de un glosario que permita aumentar la comunicación entre los expertos en el dominio del problema y los expertos en minería de datos? ¿Cuál es la relación costo beneficio del proceso de minería de datos? ¿Es rentable?

Determinación de las metas de la minería de datos*:* consiste en una traducción de los objetivos del proyecto en términos de tecnología de minería de datos.

Producción de un plan del proyecto: finalmente, se crea un plan para el proyecto que describa los pasos a seguir y las técnicas empleadas en cada paso.

Fase II

Comprensión de datos

Luego de haber establecido el problema a resolver y de haber creado un plan para hacerlo, puede centrarse en el aspecto principal de minería de datos. Las actividades a desarrollar en esta fase son:

Recolectar los datos iniciales: El primer paso es la adquisición de los datos iniciales y su preparación para futuro procesamiento. El proceso de adquisición de datos y métodos a usar para su adquisición, problemas y soluciones relacionados a la adquisición de datos.

Descripción de los datos: luego de adquirirlos, estos deben ser descritos, lo cual significa principalmente establecer el volumen de datos (numero de registros y campos por cada uno), identificación y significado de cada campo y la descripción del formato inicial de los datos.

Exploración de los datos: este paso no es obligatorio, pero si útil en muchos aspectos. El rol principal de la exploración de datos en ésta fase es encontrar una estructura general para los datos. La exploración no está directamente relacionada con la solución al problema, sino que envuelve la aplicación de pruebas estadísticas básicas que revelen propiedades en los datos recién adquiridos: si tiene campos nominales, se crean tablas de frecuencia y para los campos numéricos, se grafica su distribución y se buscan dependencias.

Verificación de la calidad de los datos: aquí se realizan chequeos sobre los datos para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los datos faltantes, encontrar valores fuera de rango (que pueden representar ruido o nuevo e interesante fenómeno): La idea en este punto es asegurar la corrección de los datos. Completitud se refiere al descubrimiento de valores erróneos en los datos y su posible solución.

Preparación de los Datos

Aunque el núcleo del proceso es la aplicación de las técnicas de modelación de la minería de datos y la evaluación de los modelos resultantes en base a sus valores predictivos o descriptivos, no debe disminuirse la importancia que tienen los esfuerzos en la preparación de los datos. La fase de la preparación de los datos está dividida en:

Selección de datos*:* un subconjunto de los datos adquiridos en las fases previas es seleccionado basado en criterio también establecido en fases anteriores: calidad de los datos que están relacionados con las técnicas de minería de datos preseleccionadas.

Limpieza de los Datos*:* este paso complementa lo anterior, también es uno de los que más tiempo consumen, debido a la enorme cantidad de técnicas que pueden aplicarse para optimizar la calidad de los datos con visitas a la fase de modelación. Algunas técnicas con: normalización de los datos (por ejemplo, de una escala decimal al rango), discreción de campos numéricos, tratamiento de valores ausentes (hay una gran cantidad, de técnicas para realizar esta tarea: reemplazo de el valor faltante con una constante global, reemplazo del valor faltante con la media, de la clase en incluso técnicas más complejas que pretenden *predecir* el valor), reducción del volumen de datos (por ejemplo, eliminando campos con bajo potencial de predicción o redundantes).

Construcción de nuevos datos: aquí se crean nuevas estructuras a partir de los datos seleccionados, por ejemplo: generación de nuevos campos a partir de dos o más ya existentes, creación de nuevos registros (muestras), fusión de dos tablas que contengan atributos diferentes para el mismo objeto, agregación de nuevos campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.

Formateo de datos*:* este paso en la preparación de los datos, implica transformaciones sintácticas de los datos sin modificar su significado, esto con la idea de permitir o facilitar el empleo de alguna técnica de minería de datos en particular.

Algunos ejemplos son: Reordenamiento de los campos y/o registros de la tabla (algunas herramientas de modelación requieren que los campos estén en cierto orden, las redes neuronales requieren que los registros estén ubicados aleatoriamente), ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (remover comas, tabulares, caracteres especiales máximos y mínimos para las cadenas de caracteres, etc.).

Fase III.

Modelación

Lo novedoso y abundante de las técnicas disponibles y de los algoritmos involucrados en la fase de modelación hacen de esta la fase más interesante del proceso de minería de datos los pasos importantes en la fase de modelación son:

Selección de la técnica de modelación: al principio del proceso se establece el problema a resolver y la meta implicada, ahora es el momento de seleccionar una técnica. Cuando se escoge una técnica se debe tener en cuenta el objetivo principal del proyecto y su relación con la principal división de las herramientas de minería de datos de acuerdo al tipo de problema. La primera división de las técnicas de modelación está hecha en base al tipo de tarea del conocimiento que sea desea:

Predicción o descripción: La tabla 1 muestra algunas clases de tareas de modelación y las técnicas de minería de datos adecuadas.

Tabla 1. Tareas de Modelación y Técnicas.

Clasificación	Métodos de inducción de reglas, árboles de decisión, k vecinos más cercanos razonamiento basado en casos.
Predicción	Análisis de regresión. árboles de regresión, redes neuronales k vecinos más cercanos
Análisis de Dependencia	Análisis de correlación, análisis de regresión, reglas de asociación, redes Bayesianas, programación con lógica inductiva.
Segmentación o Agrupación	Técnicas de agrupación, redes neuronales, técnicas de visualización.

Generación de pruebas para el modelo*:* luego de construir un modelo, se debe generar un procedimiento o mecanismo para probar la calidad y validez del modelo. Por ejemplo, en una tarea supervisada como la clasificación, es común usar la tasa de error como medida de la calidad. En consecuencia, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.

Construcción del modelo*:* una vez que la técnica de modelación ha sido seleccionada, se procede a ejecutarla sobre los datos previamente preparados para generar un modelo. Todas las técnicas de modelación tienen un conjunto de parámetros que determinan las características del modelo a general. La selección de los parámetros óptimos para la técnica de modelación, es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.

Calificación de modelo: una vez que los modelos son generados, estos son interpretados de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto y expertos en minería de datos aplican sus propios criterios (seguridad del conjunto de prueba, pérdida o ganancia de tablas, etc.).

Fase. IV

Evaluación de los resultados.

En esta fase involucra la fase de evaluación del modeló con respecto a los objetivos del proyecto y la revisión del proceso de evaluación del modeló con respecto a los objetivos del proyecto. Se debe decidir si hay o no razones para construir un modelo, deficiente (relación costo-beneficio), si es aconsejable probar el modelo en un problema real. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable calificar el modelo con relación a otros objetivos diferentes a los originales?, esto podría revelar información adicional.

Revisión del proceso: Se refiere a calificar el proceso entero de minería de datos con la idea de identificar elementos que pudieran ser mejorados.

Futuras fases. Si se han determinado que las fases hasta este momento han generado resultados satisfactorios, podrían decidirse pasar a la fase de despliegue de resultados, si no, podría decidirse por otra iteración desde la fase de preparación de datos o de modelo con otros parámetros. Podría ser incluso que en esta fase decida partir desde cero con un nuevo proyecto de minería de datos.

Fase V

Despliegue de resultados

En esta fase se define una estrategia para desplegar los resultados del proceso: monitoreo y mantenimiento: si los modelos resultantes del proceso son desplegados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias del monitoreo y mantenimiento para ser construidas sobre los modelos. La retroalimentación general por el monitoreo y mantenimiento, puede indicar si el modelo está siendo utilizado apropiadamente.

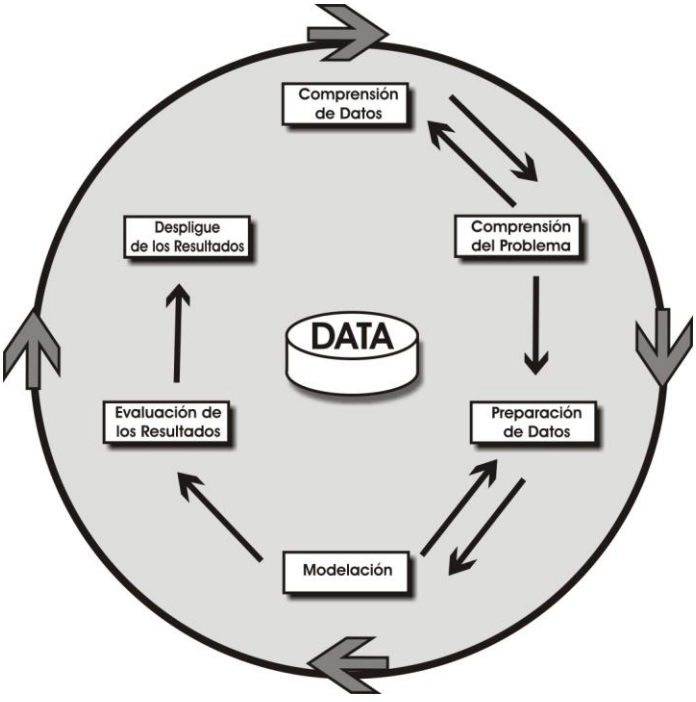

Figura 2. Proceso CRISP-dm.

Fase VI

Reporte final

Es la conclusión del proyecto. Resume los puntos importantes, la experiencia ganada y explica los resultados producidos (GAMBERGER 2001)

5.2.7 Proceso de los modelos de minería de datos

Es un modelo presentado por two crows corporation también llamado minería de datos para el descubrimiento de conocimiento, el cual toma muchas cosas de su propia experiencia, es también un proceso interactivo.

Las fases de este proceso son:

I. Definición del problema de negocios

El primero y más importante de los prerrequisitos para el descubrimiento del conocimiento es comprender los datos y al negocio. Sin esta comprensión, ningún algoritmo, por sofisticado que sea, proveerá resultados confiables. Sin este conocimiento, no será posible identificar el problema que se desea resolver, ni preparar los datos para las técnicas de minería de datos o interpretar correctamente los resultados. Deben tenerse claros los objetivos.

II. Construcción de la base de datos para la minería de datos

Esta fase y las siguientes tres son el corazón de la preparación de los datos, juntas incluyen la mayor parte del tiempo.

Los datos deben de estar reunidos en una base de datos. Nótese que esto no necesariamente implica el uso de un sistema gestor de base de datos. Dependiendo de la cantidad de datos, de su complejidad y del uso que se les pretenda dar, un archivo plano o una hoja de cálculo podría ser suficiente.

Es recomendable crear un nuevo almacén para los datos a ser analizados. Las técnicas de minería de datos producirán abundantes y frecuentes accesos al almacén de datos de la empresa, lo que podría generar problema.

Para crear esta nueva base de datos se deben realizar actividades como:

- **Recolección de datos:** se identifica la fuente de los datos que se va a emplear en el análisis. Posiblemente, los datos necesarios nunca han sido recolectados o se necesitan datos externos de base de datos, públicas (tales como censo o clima) o privadas (tales como datos sobre uso de tarjetas de crédito). Un reporte de recolección de datos muestra las propiedades de las diferentes fuentes de datos como por ejemplo: fuente (interna o vendedor externo), propietario, persona u organización responsable de su mantenimiento, costo, estructura de almacenamiento, tamaño en tablas, en registros en bytes, soporte físico del almacenamiento (CD-ROM, cinta, servidor, etc.), requerimientos de seguridad y restricciones de uso.
- **Descripción de los datos:** aquí se describe el contenido de cada archivo o tabla de la base de datos. Entre las propiedades que se enumeran en el reporte de descripción de datos están: nombre de la tabla, número de campos, numero/porcentaje de registros con datos ausentes, nombre de los campos y para cada campo el tipo de dato, definición, descripción, fuente del campo, unidad de medida, lista de valores, rango de valores, número/porcentaje de valores ausentes frecuencia, relación con la clave primaria o foránea.
- **Selección:** aquí se relaciona un subconjunto de los datos para ser procesados. No se refiere a una toma de muestra o a escoger variables predictivas, es más bien a una simple eliminación de datos irrelevantes o innecesarios. Otros criterios para excluir datos pueden ser restricciones en uso, costos, o problemas de calidad de estos.
- **Evaluación de la calidad de los datos y filtrado**: para obtener un buen modelo se necesitan buenos datos. Una evaluación de la calidad de los datos identifican características de los datos que afectarán la calidad del modelo. En esencia se pretende asegurar no sólo la correctitud y consistencia de los datos, sino también que estos correspondan a mediciones del mismo fenómeno en la misma manera. Hay varios tipos de problemas de calidad de datos como valores incorrectos, valores correctos colocados en lugares equivocados o valores ausentes.

- **Integración y consolidación:** es ahora cuando se toman los datos de distintas fuentes y se crea una única base de datos para alojarlos, lo cual requiere reconciliar las diferencias entre distintos valores de datos de varias fuentes.

Hacer esto de manera deficiente, es una de las mayores fuentes de problemas de calidad de datos.

- **Construcción de metadatos:** en esencia es crear una base de datos sobre la base de datos. Está basado principalmente en los reportes de descripción de los datos y provee información que será usada en la creación física de la base de datos y que servirá también para analizar los datos y comprender los modelos generados.

- **Carga de la base de datos para minería de datos:** en la mayoría de los casos los datos son analizados, deben estar ubicados en una base de datos independiente, y dependiendo de su cantidad y complejidad, puede ser complicado y requerir la participación de expertos.
- **Mantenimiento de la base de datos para la minería de datos:** una vez creada la base de datos necesita cuidados, debe ser respaldada periódicamente, su desempeño debe ser monitoreado, y ocasionalmente, requiere ajustes que exigen espacio de almacenamiento o aumento del rendimiento. Estas tareas involucran a profesionales en sistemas de información.

- **Exploración de datos**

  En esta fase se emplean numerosas técnicas de visualización de datos, de búsqueda de relaciones entre variables y otras medidas para exploración de los datos. La meta es identificar los campos con mayor potencial predictivo y cuales variables con valores derivados pueden ser útiles. Bases de datos con muchos campos hacen que esta tarea se ardua y consuman tiempo.

Preparación de los datos para modelación

Este es el último paso de la preparación de lo datos antes de construir modelos. Hay cuatro partes importantes en esta fase:

Selección de variables: en un caso ideal se toman todas las variables disponibles, alimentando con ellas a los algoritmos de minería de datos y dejándolos encontrar las mejores predicciones. En la práctica esto no funciona bien. Una de las razones es que el tiempo empleado para construir un modelo incrementa con la cantidad de variables.

Otra razón es que incluir variables ciegamente puede llegar a la creación de modelos erróneos.

Un error muy típico es emplear una variable de predicción que solo puede ser conocida si se conoce el valor de la variable a predecir, por ejemplo, algunas veces se ha incluido inadvertidamente la fecha de nacimiento en un algoritmo que pretende predecir la edad de un grupo de personas. Aunque muchos algoritmos de minería de datos ignoran los campos irrelevantes, no es bueno depender para todo de los algoritmos. Incluir el número de cédula como variable predictiva es en el mejor de los casos inútiles y en el pero puede reducir drásticamente la cantidad del modelo.

Selección de registros: igual que en el caso anterior, sería ideal poder emplear todos los datos disponibles, pero esto puede tardar demasiado tiempo requerir un hardware más potente por lo tanto, es buena idea tomar muestras de datos cuando la base de datos es demasiado grande. En la mayoría de los casos no habrá pérdida de calidad pero se debe de estar seguro de haber tomado las muestras al azar.

Construcción de nuevas variables: con frecuencia es necesario crear nuevas variables predictivas derivadas de datos en bruto. Por ejemplo, en la predicción del riesgo de crédito se usa más a menudo la taza deuda-ingresos como variable predictiva que los valores individuales de deuda e ingresos.

Transformación de variables: la tarea de minería de datos escogida puede dictar el cómo se presentan los datos. Por ejemplo, los campos pueden ser ajustados para que encuentren en un rango particular (en muchos casos, el rango) (0,1) muchos árboles de decisión empleados para clasificar requieren que variables continuas sean sustituidas por clases como “alto, medio, bajo”.

Construcción de modelos.

La construcción de modelos es un proceso no muy complejo. Será necesario explotar múltiples modelos alternativos hasta encontrar el más útil a la organización.

Lo aprendido durante la creación de modelos puede llevar a modificar nuevamente los datos incluso a cambiar el objetivo del proyecto. Una vez que se ha decidido que tarea de minería de datos va a efectuarse.

Se debe escoger un tipo de modelo para representar los resultados (como un árbol de decisión una red neural, algún otro modelo propietario, etc.). La selección del tipo de modelo tendrá influencia en la preparación de datos. Una vez listos los datos se pueden realizar el modelo.

La creación de modelos predictivos requiere un protocolo de entrenamiento y validación bien definido este tipo de protocolo es llamado a veces entrenamiento supervisado. En esencia, consiste en entrenar al modelo con una porción de los datos y validarlo luego con otra porción. De no hacerse de esta manera, se pueden obtener modelos sobreestimados que solo pueden predecir correctamente para el conjunto de datos procesados.

La variación se puede hacer de diversas maneras. La variación simple conste en entrenar al modelo con una porción grande de datos y dejar una más pequeña (5% al 30%) por validar. Si se dispone de un conjunto de datos pequeños, se puede hacer una variación cruzada, que consiste en separar los datos en dos grupos, entrenar a los modelos como el primer grupo y validarlo con el segundo, luego se entrena al modelo con el segundo grupo y se valida con el primero, por último se entrena al modelo con todos los datos y se usa en promedio de los errores de los modelos anteriores. De hecho el modelo más usado es el de la variación cruzada con N grupos, que sigue el mismo procedimiento para más de dos grupos.

Basados en los resultado de la construcción de modelo, se puede decidir sin crear otros modelos empleados, la misma técnica con parámetros diferentes o intentar con otra técnica o algoritmo.

Evaluación del modelo

Después de construir un modelo se debe evaluar e interpretar sus resultados. Debe tenerse presente que la confiabilidad calculada para el modelo solo aplica para los datos sobre los que se ha realizado el análisis se pueden emplear múltiples medios para interpretación de resultados. Las matrices de conjunción son muy empleadas de problemas de clasificación.

Despliegue de modelos y resultados

Una vez que el modelo ha sido construido y validado puede ser empleado en una o dos maneras importantes: la primera de ellas consiste en que un analista recomiende acciones basadas simplemente en la observación del modelo y sus resultados. La segunda consiste en aplicar el modelo a diferentes conjuntos de datos, por ejemplo, para marcar ciertos registros según su clasificación o asignarles puntuación tales como la probabilidad de acción.

A veces los modelos por parte del proceso de negocios. Tales como análisis de riesgo, autorizaciones de crédito y detecciones de fraude, en esos casos el modelo es incorporado a una aplicación de negocios (EDELSTEIN, 1999).

5.3 RELACIONES CON OTRAS TÉCNICAS DE CONOCIMIENTO

5.3.1 Datawerehouse

Actualmente, cada vez más empresas se encuentran en la tarea de integración de todos los procesos con el objetivo de tomar decisiones en forma rápida y acertada de acuerdo a las necesidades cambiantes de su mercado, para esto es necesario tener el soporte de una tecnología que les ofrezca la disponibilidad de su información en una manera eficiente, confiable y en un tiempo de respuesta corto, una opción es la implementación y uso de un data werehouse.

El objetivo del data werehouse es hacer llegar la información correcta en el tiempo correcto a los tomadores de decisiones (BERRY, 2000). La idea detrás de un data werehouse es acumular todos los datos de una compañía en una sola fuente de datos lógica para brindar una mayor visibilidad de los procesos de negocios, generar conocimiento y mejorar el desempeño organizacional. (KALALOTA Y ROBISON, 2000).

En la mayoría de las organizaciones se pueden encontrar bases de datos extensas en operación con el objetivo de dar el soporte a las transacciones diarias de la empresa.

Este tipo de base de datos son conocidas como bases de datos operacionales o transaccionales; en la mayoría de los casos estas no han sido designadas para almacenr datos históricos ya que solo tienen el objetivo de soportar todas las transacciones de los sistemas operaciones. El segundo tipo de bases de datos encontrado dentro de las organizaciones son las data werehouses, las cuales han sido designadas para el soporte a las decisiones estratégicas, y son generalmente construidas a partir de las bases de datos operacionales. La característica básica de un data werehouse es que contiene una vasta cantidad de datos los cuales están destinados a hacer realizados para obtener tendencias, áreas de oportunidad con información histórica. (ADRIAANS Y ZANTINGE, 1997)

El concepto de data werehouse incluye conceptos comunes en base de datos, sin embargo, la diferencia radica en el enfoque, ya que una base de datos tiene propósitos operacionales y transaccionales mientras data werehouse tiene propósitos analíticos y orientados al soporte en las tomas de decisiones (CASTAÑEDA, 2001).

Existen diversas definiciones para el data werehouse cada una con diferentes enfoques pero todas esquematizan los mismos elementos a continuación se listan algunas definiciones:

“Sistema computacional diseñado para proporcionar un acceso instantáneo a la información por parte de los tomadores de decisiones. El data werehouse copia sus datos desde sistemas existentes como entradas de orden, recursos humanos, etc. Y los almacena para uso ejecutivo en lugar de programadores. Los usuarios del data warehouse utilizan un software especial que permite crear y accesar la información que necesitan” (DATABASE, INC, 2000).

“Es un deposito o almacén de datos actuales e históricos en donde se almacena información que es solicitada y necesitada por los tomadores de decisiones la cual es utilizada para manejar sus negocios correctamente , predecir, y hacer frente a los cambios que se presentan día a día. Este repositorios contienen datos internos de la empresa, datos externos así como información del mercado, industria entre otros” (MEJIA, 1996).

La definición elástica de data warehouse es la siguiente la cual pertenece al padre del data warehouse bill innmon y dice:

“El data warehouse es una colección de datos orientados al tema integrados, no volátiles e historiados, organizados para el apoyo de un proceso de ayuda a la decisión”(MICROSOFT, 1996).

De los anterior se puede concluir que un data werehouse es un proceso de adquisición, consolidación, almacenamiento y administración de toda la información de los procesos de una empresa alineados a la estructura del negocio con el objetivo de proporcionar a los tomadores de las decisiones información relevante que apoye a la estrategia del negocio.

Data warehouse debe permitir al negocio ser proactivo en su mercado, es decir, decidir y anticipar en función de la información disponible y capitalizar sobre sus experiencias la información es vital para los negocios. Todos los datos, deben organizarse, coordinarse integrarse y almacenarse para dar al usuario una visión orientada a su negocio y sin duda es data warehouse lo que lleva acabo.

En un data warehouse además de tomar decisiones operativa basadas en valores habituales, otros usuarios podrían trabajar sobre historiales, lo que permitiría decisiones a largo plazo, posicionamientos estratégicos y análisis de tendencias. Es una especie de punto focal que guarda en un único lugar toda la información útil proveniente de sistemas de producción y de fuentes externas. (INMOM, 1994).

En resumen data warehouse es el sistema para el funcionamiento y distribución de cantidades masivas de datos.

5.3.1.1 Características de data warehouse

Según (INOM, 1994), un data warehouse debe de cumplir con las siguientes características:

- Integrado: los datos almacenados en data warehouse deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las diversas necesidades de los usuarios.
- Temático: sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales. Por ejemplo, todo los datos sobre clientes pueden ser consolidados en una única tabla del data warehouse. De esta forma las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información esta contenida en el mismo lugar.
- Histórico: el tiempo es parte implícita de la información contenida en un data warehouse. En los sistemas operacionales los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en el data warehouse sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el data warehouse se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones y hacer pronósticos.

- No volátil: el almacén de información de un data warehouse para ser leído, y no modificado la información es por tanto permanente, significando la actualización del data warehouse la incorporación de los últimos valores que tomaron las distintas variables contenidas en el sin ningún tipo de acción sobre lo que ya existía.

5.3.1.2 Procesos del data warehouse

Para comprender el concepto del data warehouse, es importante considerar los procesos que lo conforman, en la figura 3 se muestra los procesos claves en la gestión de un Data Warehouse. (SANCHEZ Y CRIADO, 2002).

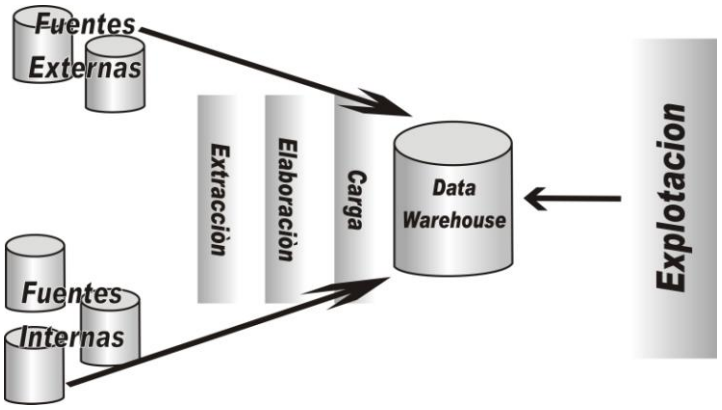

Figura 3. Procesos del Datawarehouse.

- Extracción: obtención de información de las distintas fuentes tanto internas como externas.
- Elaboración: filtrado, limpieza, depuración, homogeneización y agrupación de la información.
- Carga: organización y actualización de los datos y los metadatos en la base de datos.
- Explotación: extracción y análisis de la información en los distintos niveles de agrupación.

MEJÌA (1996), menciona que un data warehouse es efectivo para una empresa, debe cubrir los siguientes puntos:

- Debe ser capaz de monitorear las operaciones actuales del negocio y compararlas con las operaciones hechas en el pasado.
- Debe ser capaz de hacer pronósticos de operaciones futuras de una manera racional, encontrando nuevos procesos del negocio con lo que se puede producir nuevas operaciones que apoyen a dichos procesos.

- Debe ser capaz de encontrar relaciones entre los datos producidos en las diferentes aéreas de la empresa, encontrando patrones que permitan predecir futuros sucesos.

5.3.1.3 Alternativas para aplicar un data warehouse

La lista siguiente muestra alternativas en las que puede aplicarse un data warehouse:

- Un almacenamiento central único en el que todas las consultas de la organización son ejecutadas.
- Accesibilidad a los datos para la mayoría de las herramientas y plataformas. Entre estas herramientas se encuentran las de bajo nivel, como una consulta simple en muchas hojas de trabajo o herramientas de análisis multidimensionales.
- Acceso a reportes y consultas estándares. Para proporcionar esta característica, es necesario un software de servidor de reportes (incluyendo los de interfaces de web).
- Sirven de apoyo para las nuevas ciencias como data mining.
- El data warehouse puede ser utilizado por otras aplicaciones como la fuente de sistemas operaciones de datos, pueden alimentar datos a otros data warehouse (GUPTA, 1997).
- Entrega de información inmediata, debido a que data warehouse compactan el tiempo entre la ocurrencia de eventos del negocio y la alerta al ejecutivo.
- Integración de datos externos e internos de la organización.
- El data warehouse probé una fotografía completa de las compañías.
- Visión futura basándole en tendencias históricas. Los data warehouse contiene datos de muchos años.

- Herramientas de visualización de datos de nuevas maneras.

- Libertad de las limitaciones del departamento de sistemas de información. Uno de los problemas de los sistemas computacionales es que requieren expertos para utilizarlos con el data warehouse ya no es necesario que pasen días enteros para que se genere un reporte ya que proporciona a los usuarios la facilidad de general su propio reporte (DATA BASE INC; 2000).

5.3.1.4 Metodología de implementación de un data warehouse

Tal y como lo menciona Johnson (1988). En un artículo de computer Word: "Un data warehouse no se puede comprar, se tiene que construir". La construcción e implementación de un data warehouse es un proceso evolutivo. Este proceso se tiene que apoyar en una metodología específica para este tipo de procesos SAS institute propone la metodología "Rapid warehusing methodology" (SAS INSTITUTED,2001).

Dicha metodología es iterativa, y está basada en el desarrollo incremental del proyecto del data werehouse, la figura 4 muestra las cinco fases de esta metodología así como sus relaciones. A continuación se explica cada una de las fases.

1.-Definición de objetivos.- Abarca los siguientes puntos: que tipo de empresa se implantará el data werehouse, el objetivo general y especifico del data werehouse.

2. Definición de los requerimientos de información.- Este punto ayudará a justificar la asignación de recursos hacia la construcción del data werehouse. Algunas de las preguntas que deberán ser contestadas: ¿Quién será el usuario final?, ¿Qué herramienta del usuario final será usada?, ¿Qué plataformas son utilizadas actualmente o contempladas en el futuro?, ¿Dónde esta almacenada la base de datos original?, ¿Cuando se necesitará que el data warehouse este en operación? (FUTURETEANALYTICS, 1988).

3. Diseño y modelización.- Los requerimientos de información identificados durante la anterior fase proporcionaron las bases para realizar el diseño y modelización del data warehouse. En esta fase se identificarán las fuentes de datos (sistema operacional, fuentes externas etc.) y las transformaciones necesarias para, a partir de dichas fuentes, obtener el modelo lógico de datos del data warehouse.

Este modelo estará formado por entidades y relaciones que permitirán resolver las necesidades de negocios de la organización. El modelo lógico se traducirá posteriormente en el modelo físico de datos que se almacenará en el data warehouse y que definirá la arquitectura de almacenamiento data werehouse adaptándose el tipo de explotación que se realiza del mismo. La mayor parte de estas definiciones de los datos estarán almacenadas en los metadatos y formaran parte del mismo.

4. Implementación.- Esta fase lleva implícitos los siguientes pasos.

- Extracción de los datos del sistema operacional y transformación de los mismos
- Carga de los datos validados del data warehouse. Esta carga deberá ser planificada con una periodicidad que se adaptara a las necesidades detectadas durante las fases de diseño del nuevo sistema.
- Explotación del data warehouse mediante diversas técnicas dependiendo del tipo de aplicación que se aplique a los datos:
- La información necesaria para mantener el control sobre los datos se almacena en los metadatos técnicos (describen las características físicas de los datos) y de negocio (describen como se usan esos datos). Con la finalización de esta fase se obtendrá un data warehouse disponible para su uso por parte de los usuarios finales y el departamento de informática.

5. Revisión.- La construcción del data warehouse no finaliza con la implantación del mismo, sino que es una tarea iterativa en la que se trata de incrementar su alcance aprendiendo de las experiencias anteriores. Después de implantarse, debería realizarse una revisión del data warehouse planteando preguntas que permitan después de los 6 o 9 meses posteriores a su puesta en marcha, definir cuales serian los aspectos a mejorar o potenciar en función de la utilización que se haga del nuevo sistema una fase opcional es el diseño de la estructura de cursos de formación. Con la información obtenida en reuniones con los distintos usuarios es posible realizar una serie de cursos a la medida que tiene como objetivo el proporcionar la información estadística necesaria para el mejor aprovechamiento de la funcionalidad incluida en la aplicación. Para posteriormente realizar prácticas sobre el desarrollo realizado, las cuales permitirán fijar los conceptos adquiridos y servirán como formación a los usuarios.

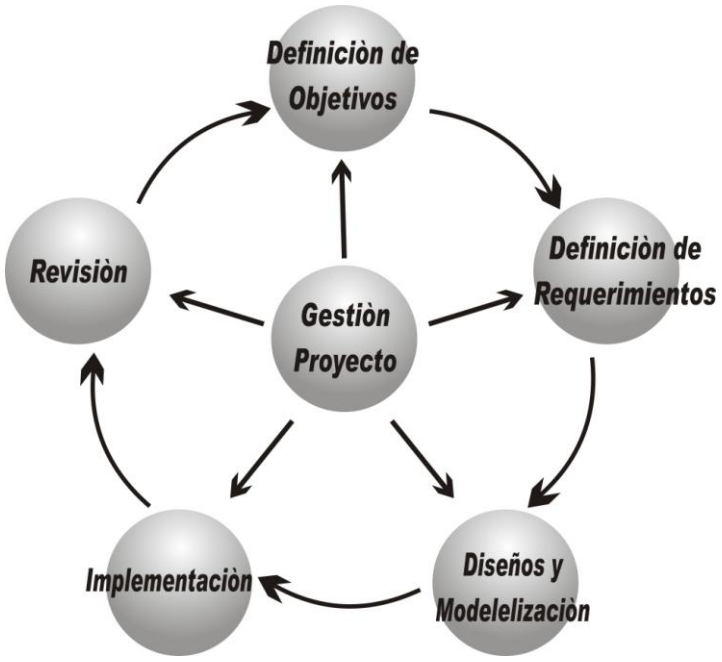

Figura 4. Fases de implementación de un Data Warehouse

5.3.1.5 Elementos de una arquitectura data warehouse

Una de las razones por las que el desarrollo de un data warehouse crece rápidamente es que realmente, es una tecnología muy entendible una arquitectura data warehouse es una forma de representar la estructura total de datos, comunicación, procesamiento, y presentación, que existe para los usuarios finales que disponen de una computadora dentro de la empresa.

La arquitectura se constituye de un número de partes interconectadas: base de datos operaciones / nivel de base de datos externo nivel de acceso a la información nivel de acceso a los datos, nivel de directorio de datos. (Meta data) nivel de gestión de proceso nivel de mensaje de la aplicación nivel de data warehouse nivel de organización de datos.

5.3.2 Base de datos operacional /Nivel de datos externo

Los sistemas operacionales procesan datos para apoyar las necesidades operacionales críticas. Pára hacer eso, se han creado las bases de datos operacionales históricas que proveen una estructura de procesamiento eficiente, para un número relativamente pequeño de transacciones comerciales bien definidas.
Sin embargo, a causa del enfoque limitado de los sistemas operacionales, las bases de datos diseñadas para soportar estos sistemas, tiene dificultad al accesar a los datos para otra gestión o propósitos informáticos. Esta dificulta en acceder a los datos operacionales es amplificar por el hecho que muchos de estos sistemas tienen de 10 a 15 años de antigüedad. El tiempo de algunos de estos sistemas significa que la tecnología de acceso a los datos disponibles para obtener los datos operacionales, es así mismo antigua. Ciertamente, la meta del data warehousing es liberar la información que es almacenada en base a datos operacionales y combinarla con la información desde otra fuente de datos, generalmente externa.

Cada vez más, las organizaciones grandes adquieren datos adicionales desde bases de datos externas. Esta información incluye tendencias demográficas, adquisitivas y competitivas (que pueden ser proporcionadas por instituciones oficiales INEI). Internet también llamada "información superhighway" (supercarretera de la información) provee el acceso a mas recursos de datos todos los días.

5.3.2.1 Nivel de acceso a la información

El nivel de acceso a la información de la arquitectura data warehouse, es el nivel del que el usuario final se encargara directamente. En particular, representa las herramientas que el usuario final normalmente usa día a día. Por ejemplo: Excel, Lotus1-2-3, Focus, Acces, SAS, etc.

Este nivel también incluye el hardware y software involucrados en mostrar información en pantalla y emitir reporte de impresión, hojas de cálculo, gráficos y diagramas para el análisis y presentación.

Hace dos décadas que el nivel de acceso a la información se ha expandido enormemente, especialmente a los usuarios finales quienes se han volcado a las PCs monousuarias y las PCs en redes.

Actualmente, existen herramientas más y mas sofisticadas para manipular, análisis y presentar los datos sin embargo, hay problemas significativos al tratar de convertir los datos tal como han sido recolectados y que se encuentran contenidos en los sistemas operacionales en información fácil y transparente para las herramientas de los usuarios finales. Una de las claves para esto es encontrar un lenguaje de datos común que puede usarse a través de toda la empresa.

5.3.2.2 Nivel de acceso a los datos

El nivel de acceso a los datos de la arquitectura data warehouse está involucrado con el nivel de acceso a la información para conversar en el nivel operacional. En la red mundial de hoy, el lenguaje de datos común que ha surgido es SQL. Originalmente SQL fue desarrollado por IBM como un lenguaje de consulta, pero en los últimos veinte años ha llegado a ser el estándar para el intercambio de datos.

Uno de los adelantos claves de los últimos años ha sido el desarrollo de una serie de "filtros" de acceso a datos, tales como EDA/SQL para acceder a casi todos los sistemas de gestión de base de datos (Data Base Management Systems- DBMSs) y sistemas de archivos de datos, relacionales o no. Estos filtros permiten a las herramientas de acceso a la información, acceder también a la data almacenada en sistemas de gestión de base de datos que tienen veinte años de antigüedad.

El nivel de acceso a los datos no solamente conecta DBMSs diferentes y sistemas de archivos sobre el mismo hardware, sino también a los fabricantes y protocolos de red. Una de las claves de una estrategia data warehousing es proveer a los usuarios finales con "acceso a datos universales".

El acceso a los datos universales significa que, teóricamente por lo menos, los usuarios finales sin tener en cuenta la herramienta de acceso a la información o ubicación, deberían de ser capaces de acceder a cualquier o a todos los datos en la empresa que es necesaria para ellos, para hacer su trabajo.

El nivel de acceso a los datos entonces es responsable de la interface entre las herramientas de acceso a la información y a las bases operacionales. En algunos casos, esto es todo lo que un usuario final necesita. Sin embargo, en general las organizaciones desarrollan un plan mucho mas sofisticado para el soporte de data werehousing.

5.3.2.3 Nivel de dirección de dato (Metadata).

A fin de proveer el acceso a los datos universales, es absolutamente necesario mantener alguna forma de directorio de datos o repositorio de la información metadata.

La metadata, es la información alrededor de los datos dentro de la empresa. Las descripciones de registro en un programa COBOL son metadata. También los son las sentencias DIMENSION en un programa FORTRAN o las sentencias a crear SQL a fin de tener un depósito totalmente funcional, es necesario tener una variedad de metadata disponibles, información sobre las vistas de datos de los usuarios finales e información sobre las bases de datos operacionales. Idealmente, los usuarios finales deberían accesar a los datos desde el data warehouse (o desde la base de datos operacionales), sin tener que conocer donde residen los datos o la forma en que se han almacenado.

5.3.2.4 Nivel de gestión de procesos

El nivel de gestión de procesos, tiene que ver con la programación de diversas tareas que deben realizarse para construir y mantener el data warehouse y la información del directorio de datos. Este nivel puede detener el alto nivel de control de trabajo para muchos procesos (procedimientos) que deben de ocurrir para mantener el data warehouse actualizado.

5.3.2.5 Nivel de mensaje de la aplicación

El nivel de mensaje de la aplicación tiene que ver con el transporte de información alrededor de la red de la empresa. El mensaje de aplicación se refiere también como “subproducto”, pero puede involucrar sólo protocolos de red.

Puede usarse por ejemplo, para aislar aplicaciones operacionales o estratégicas a partir del formato de datos exacto, recolectar transacciones o los mensajes y entregarlos a una ubicación segura en un tiempo seguro.

5.3.2.6 Nivel data warehouse (físico)

En el data warehouse (núcleo), es donde ocurre la data actual, usada principalmente para uso estratégico. En algunos casos, uno puede pensar data warehouse simplemente como una visita lógica o virtual de datos. En muchos ejemplos data warehouse puede no involucrar almacenamiento de datos. En un data warehouse físico, copias, en algunos casos, muchas copias de datos operaciones y/o externos, son almacenados realmente de una forma que es fácil de accesar y es altamente flexible. Cada vez más, los data warehouse son almacenados sobre plataformas cliente/servidor, por lo general se almacenan main frames.

5.3.2.7 Nivel de organización de datos

El componente final de la arquitectura de data warehouse, es la organización de los datos. Se llaman también gestión de copia o de réplica, pero de hecho, incluye todos los procesos necesarios como seleccionar, editar, resumir, combinar, y cargar datos en el depósito y accesar a la información desde bases de datos operaciones y/o externas.

La organización de datos involucra con frecuencia una programación compleja pero cada vez más, están creándose las herramientas data warehousing para ayudar en este proceso. Involucra también programas de análisis de calidad de datos y filtros que identifican modelo y estructura de datos dentro de la data operacional existente.

5.3.3. Datamarts

Desacuerdo a Johnson (1999). El data mar,t es un conjunto especializado de información de negocio enfocado a un aspecto particular de la empresa, como los departamentos y lo procesos de negocios, la información en un data mart siempre viene de muchos sistemas de datos.

Muchas compañías deciden alimentar a los data maris desde los data warehouses debido a que la información en el data warehouse ya está consolidada y procesada desde la misma fuente. La estrategia de data mart apareció para ser mas popular y fácil de entender la creación de un data mart orientada a un área para resolver problemas particulares representan una solución más simple.

5.3.4. Almacenamiento de datos operacionales

Es una variación de un sistema de procesamiento de transacciones en línea (OLTP,On Line Transactianl Procesing), o mejor conocidos como los sistemas operacionales.

La diferencia, con los operacionales, radica en que contiene un sistema hibrido OLTP y un sistema analítico. Contiene información que es frecuentemente actualizada de acuerdo a bases personales, y con el propósito de responder a cambios en los sistemas operacionales comunes, opuesto a las actualizaciones periódicas a un data warehouse.
Los datos son extraídos de los sistemas operacionales, transformados y agregados de acuerdo a una limitante o preformado específico. El propósito el propósito es proveer un sistema de consulta a nivel operacional que no afecte el desempeño de los sistemas operacionales (GUPTA, 1997).

5.3.5 Olap

Por medio de esta tecnología el usuario puede elaborar complejas consultas basadas en el análisis de la información desde la perspectiva de las múltiples dimensiones del negocio. Dichas dimensiones del negocio se estructura a su vez en distintos niveles de detalles (por ejemplo, la dimensión geográfica puede constar de los niveles nacional, estatal, municipal y distrital).

La funcionalidad de los sistemas OLAP se caracteriza por ser un análisis multidimensional de los datos corporativos, que soportan loa análisis del usuario y la posibilidad de “navegación”, seleccionado la información a obtener de una manera simple y grafica (MARCH Y LUNA, 2001).
En general, estos sistemas OLAP deben:

- Soportar requerimientos complejos de análisis
- Analizar datos desde diferentes perspectivas
- Soportar análisis complejos contra un volumen inmenso de datos

En 1994 Codd y Codd introdujeron 12 reglas sobre el modelo OLAP, las cuales son las siguientes:

- Vistas multidimensionales, manejo y organización conceptual y física de la información en forma multidimensional.
- Transparencia. Capacidad para acceder a datos de otras fuentes de manera sencilla y transparente.
- Accesibilidad. Habilidad para obtener información completa y estructurada, de fuentes externas de datos tales como bases de datos, archivos planos etc.
- Desempeño y consistencia. El número de dimensiones utilizadas en el sistema no debe degradar el desempeño del sistema ni tampoco afectar la consistencia de la información.
- Cliente / Servidor. Las herramientas deben de poder operar en ambiente cliente / servidor.

- Dimensionalidad genérica. Cada dimensión deberá ser tratada de igual manera.
- Uso eficiente del almacenamiento. Manejo eficiente de la "porosidad" de la base multidimensional. Para ocupar la mínima cantidad de espacio. Por "porosidad" se entiende la manera que las herramientas manejan el espacio requerido para almacenar la información, debido a que en la estructura de los datos de bases multidimensionales, se cuentan con muchas "celdas" o campos vacios. Un buen manejo de la porosidad implica que la herramienta es capaz de detectar las celdas vacías, y administrar eficientemente el espacio que estos requieren.

5.3.6. Knowledge (KDD)

El proceso de obtención de conocimiento o como lo llaman Adriaans y Zantingue (1997), descubrimiento en bases de datos (Knowledge Discovery in Databases o simplemente KDD), se encuentra formando por varias fases, siendo una de las mas utilizadas es la minería de datos, y precisamente esta tecnología es la que nos permite descubrir información oculta en los datos de cualquier sistema.

5.3.6.1. Fases KDD

A continuación se explica brevemente cada una las fases o etapas identificadas en el descubrimiento del conocimiento en las bases de datos propuesta por Fayad y Piatetsky (2000). El proceso inicia con los datos en bruto y finaliza en la extracción del conocimiento el cual se adquiere de las siguientes etapas:

- Selección. Es la selección o la segmentación de los datos de acuerdo a algún criterio. Por ejemplo: "todas las personas que posean un coche", de esta manera el subconjunto de datos es delimitado al espacio muestral de donde se desea obtener un conocimiento.
- Procesamiento. En esta etapa se limpian o depuran los datos donde cierta información es removida, por ser clasificada como innecesaria, debido a que puede afectar el tiempo de búsqueda. Por ejemplo, se considera innecesario el sexo de in paciente al detectarle un embarazo. También los datos son reconfigurados para asegurar un formato, debido a que hay posibilidad en inconsistencia de presentación ya que el dato puede venir de diferentes fuentes, por ejemplo el sexo puede ser representado con una "F" o "M" y en otros casos con "1" o "O".
- Transformación. En esta fase los datos son convertidos para que sean accesados y navegables. Los datos son transformados por medio de reglas de acuerdo el tipo de negocios.
- Minería de datos. En esta fase consiente con la extracción de los patrones de los datos. Un patrón puede estar definido dado un conjunto de hechos (datos) F. un lenguaje L, y alguna medida de certeza C.

- Interpretación y evaluación. Los patrones identificados por el sistema son interpretados en conocimiento, el cual puede ser usado para soporte en las decisiones humanas. Por ejemplo, la predicción y clasificación de tareas, sumarizacion de contenido de una base de datos o explicación de un fenómeno observado.

5.3.7. Relaciones existentes

Las herramientas OLAP, permiten navegar a través de los datos almacenados dentro del data warehouse y analizarlos dinámicamente desde una perspectiva multidimensional, es decir, considerado unas variables en relación con otras y no de forma independiente entre sí, permitiendo enfocar el análisis desde distintos puntos de vista.

Sin embargo, el análisis OLAP depende de un usuario que plantee una consulta o hipótesis. Es el usuario el que lo dirige y, por lo tanto, el análisis queda limitado por las ideas preconcebidas que aquel pueda tener. La minería de datos constituye un paso más en el análisis de los datos de la empresa para apoyar a la toma de decisiones. No se trata de una técnica que sustituye al análisis OLAP, sino que los complementa, permitiendo realizar un análisis más avanzado de los datos y extraer más información de ellos. Como

ya se ha comentado anteriormente, la minería de datos, es el propio sistema el que descubre nuevas hipótesis y relaciones, de este modo el conocimiento obtenido con estas técnicas no queda limitado por la visión que el usuario tiene del problema.

Al identificar los modelos de data mining el tomador de las decisiones, podrá seleccionar cual es que el más le conviene para una situación especificar y aplicar algún modelo al data warehouse del OLAP, el cual producirá relaciones o segmentación de datos, pudiéndose crear un data mart especifico para el problema que está resolviendo el OLAP. Las ventajas serán que un manejo de información en menor cantidad, depurada y de mas valor, mejorando los resultados del OLAP.

5.4 EVOLUCIONES.

5.4.1 Mundial

A nivel mundial son mayores las implementaciones que se han dado en el campo del data mining a continuación se presentan algunos casos en donde esta técnica ha sido implementada con éxito.

Especialmente en las economías de Latino América, Rusia, India y China, ya que estas se basan en los hechos para tomar decisiones.

En los 90´s, cuando empezó a tener más auge el data mining la British Broadcasting Corporation (BBC) el Reino Unido empleo 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 utilizaba redes neuronales y arboles de decisión aplicados a datos históricos de la cadena para determinar los criterios que participaban según el programa se pretendía analizar. La versión final se desempeño tan bien como un experto humano con la ventaja de que se adapto más fácilmente a los cambios porque era constantemente actualizada con datos actuales.

En España el estudio fue desarrollado en una empresa de telefonía española que básicamente situó sus objetivos en dos puntos: el análisis del perfil de los clientes que se dan baja y la predicción del comportamiento de sus nuevos clientes. Se analizaron los diferentes históricos de clientes que habían abandonado la compañía (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, internacionales y gratuitas.

Al contrario de lo que se podría pensar, los clientes que abandona la compañía telefónica generaban ganancias para la empresa; sin embargo, una de las conclusiones más importantes radico en el hecho de que los clientes que se daban de baja recibían pocas promociones y registraban un mayor número de incidencias respecto a los otros.

De esta forma se recomendó a la compañía hacer un estudio sobre sus ofertas y analizar profundamente las quejas y comentarios recibidos por esos clientes. Al describir el perfil que presentaban, la operadora tuvo que diseñar un trato más personalizado para sus clientes actuales con esas características. Para predecir el comportamiento de sus nuevos clientes se diseño 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.
En Italia el club Ac Milan, utilizó redes neuronales para prevenir lesiones y optimizar el acondicionamiento de cada atleta. Esto ayudó a seleccionar el fichaje de un posible Jugador o a alerta médico del equipo de una posible lesión. El sistema, creado por Computer Associates International, fue alimentado por datos de cada jugador, relacionados con su rendimiento, alimentación y respuesta a estímulos externos, que se obtenían y analizaban cada quince días. El jugador llevaba a cabo determinadas actividades que eran monitoreadas por veinticuatro sensores conectados al cuerpo y que trasmitían señales de radio que posteriormente eran almacenados en una base de datos.
Actualmente el sistema dispone de 5.000 casos registrados que permiten predecir alguna posible lesión. Con ello, el club interna ahorra dinero evitando comprar jugadores que presenten una alta posibilidad de lesión, lo que hace incluso renegociar su contrato.

Por otra parte, el sistema pretende encontrar las diferencias entre las lesiones de atletas de ambos sexos, así como saber si una determinada lesión se relaciona con el estilo de juego de un país concreto donde se practica el futbol.

En los Estados Unidos, el data mining se ha ido incorporando a la vida de las empresas, gobiernos, universidades, hospitales y diversas organizaciones que están interesadas en explorar sus bases de datos.

El Advanced Scout, es un software que emplea técnicas de data mining y que han desarrollado investigaciones 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 (doublé team) a un jugador por ejemplo contario, entre otros. El objetivo, es ayudar a los entrenadores a aislar eventos que no detectan cuando observan el juego en vivo o en grabación.

Un resultado interesante fue uno hasta entonces no observado por los entrenadores de Knicks de New York. El doble marcaje a un jugador puede generalmente dar la oportunidad a otro jugador de encestar más fácilmente.
En el 2001, las instituciones financieras a escala mundial perdieron más de 2,000 millones de dólares en fraudes con tarjetas de crédito y debito el Falcon Fraud Manager, es un sistema inteligente que examinara transacciones, propietarios (MOLINA, 2002).

5.4.2 Nacional

A nivel investigación es nulo el avance que ha tenido nuestro país debido a que el data mining en México es una herramienta muy poco empleada (FRAGOSO, 2001) es muy difícil identificar a las empresas que lo están empleando.
Se sabe de un caso en México, referente al área de bibliotecas, que consistió en la creación de dos bases que son: CIME (Ciencia en México) y CIME B (Científicos Mexicanos).

El objetivo de formar una base de datos con los artículos científicos publicados en el siglo XIX y ponerlas a disposición de Latinoamérica, es para difundir el pasado científico de nuestro país.
La base de datos CIME (Ciencia en México) fue el resultado de las inquietudes surgidas durante el Seminario Internacional: Problemas técnicos de la documentación de la historia de las ciencias y la tecnología en América Latina, que tuvo lugar en la Ciudad de México, del 21 al 25 de enero 1985.

El objetivo del control era saber lo que el país había producido en materia de documentación científica, para que las partes nacionales se integren a una red bibliográfica internacional, y una vez identificada se ofrecía al usuario a través del

programa de Disponibilidad Universal de Publicaciones (D.U.P.) de la Federación Internacional de Asociaciones de Bibliotecas y Bibliotecarios (IFLA).

La finalidad de la base es: habiendo hecho una revisión de la literatura científica aparecida en publicaciones seriadas mexicanas del siglo XIX, codificar cada uno de los artículos para finalmente poner a disposición de los investigadores toda esa información en forma automatizada.

CIME, pretendía cubrir las áreas temáticas de la agricultura, biología, farmacología, física, geografía, geología, meteorología, mineralogía, química, tecnología, zoología, entre otras.

Cada registro de CIME contenía la siguiente información: autor del artículo, tipo de documento (revista o periódico), nombre de la revista, si era traducción u original, lugar de edición, editor, volumen, fecha, páginas, ilustraciones, bibliografía, impresor, localización, publicación de origen cuando era traducción, descriptores y notas.

Las fuentes para obtener la información fueron las propias revistas del siglo XIX; de estas compilaron 140 títulos, de los cuales se analizaron 46 y de estos 198 volúmenes.

Una vez iniciados los trabajos de CIME, se detecto la ausencia de biografías completas de una gran parte de los científicos mexicanos. Haciendo un análisis de la situación, se consideró que era necesario crear una base de datos que contuvieran la labor y producción bibliográfica de todos estos personajes, además de todos sus datos personales, con el fin de difundir posteriormente la información.

Con estas dos bases de datos se puede conocer cuantas documentaciones científicas se habían hecho en el país y por consecuencia cuantos investigadores mexicanos habían publicado artículos de esta índole.

Otro caso más reciente aquí en nuestro país referente a data mining se presento en el sector educativo, se hizo un estudio sobre los recién titulados de la carrera de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Chihuahua. Se quería observar si los recién titulados se insertaban en actividades profesionales relacionadas con sus estudios y, en caso negativo, se buscaba saber el perfil que caracterizó a los ex/alumnos 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 hacían una buena inserción laboral o si existían 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 habitada el estudiante, colegio de donde provenía, nota al ingresar y promedio final al salir de la carrera. A partir de estos resultados, la universidad tubo que hacer un estudio socioeconómico sobre grupos de alumnos aquí pertenecía a las clases económicas bajas para dar posibles soluciones, debido a que tres de las cuatro variables no dependían de la universidad (rodas, 2001).

En el Laboratorio de Sistemas de Información del Centro de Información en Computación (CIC) del Instituto Politécnico Nacional (IPN), se desarrollo una herramienta que forma parte del proyecto ANASIN, con la cual la minería de datos se realiza utilizando la técnica que construye cubo de n-dimensiones conocida como generalización y sumarizacion en cubos de datos, que es implantada en una base de datos relacional. La generalización de los datos se puede desarrollar en los niveles que se considere necesario usar y así realizar análisis a diferentes niveles de conceptos. En comportamiento; al término de las búsquedas los resultados se muestran en reportes de tipo texto y gráficas.

En este proceso de minería se pueden distinguir dos tipos de programas: los que extraen la región de interés de la base de minería, llamados extractores; y los programas que realizan la búsqueda de patrones, llamados mineros. Tanto la actividad de extracción como la búsqueda de patrones puede consumir demasiado tiempo, por lo cual se delegan a programas que las realizan en formas autónoma y nocturno y así aprovechar los recursos computacionales.

La herramienta desarrollada en el laboratorio, llamado modulo de minería de datos ANASIN, tiene el modelo de trabajo cliente / servidor, donde se distinguen tres actividades básicas.

- Solicitudes de minería, realizadas en una estación de trabajo o cliente.
- El proceso de minería o generación de región y búsqueda de un patrón determinado en el servidor.
- La visualización del resultado en el cliente.

ANASIN, comprende un conjunto de herramientas y métodos para recolectar, integrar y analizar datos en una organización distribuida.

El trabajo en cuestión presentan los resultados en forma gráfica, considerado que las variables analizadas con elegida por el mismo usuario, lo cual evita presentar resultados que no sean considerados relevantes para su propósito particular.

Esta herramienta se utiliza como apoyo en varios proyectos que se han planeado, entre los cuales se pueden mencionar: uso de agentes para la minería de datos, uso de agentes para la minería distribuida, uso de agentes para la minería en texto, generación de nuevos agentes (GUZMAN, 1999).

En los últimos años el crecimiento del uso de data mining se debe a que esta herramienta es muy usada en sistemas de Business Intelligence y Administración del conocimiento, algunas herramientas que empiezan a ser utilizadas en las empresas mexicanas con el CRM (Customer Relationship Management).

5.5 TENDENCIAS EN INVESTIGACIÒN

5.5.1 Aplicaciones más frecuentes

Existen numerosas áreas donde la minería de datos se puede aplicar, prácticamente en todas las actividades humanas que generen datos.

Aquí se presenta una lista con las principales áreas en donde a tenido éxito la aplicación de minería de datos.

Comercio

- Identificar patrones de compra
- Buscar asociaciones entre clientes y características demográficas
- Predecir respuestas a campañas de marketing

Bancos

- Detectar patrones de uso fraudulento de tarjetas
- Identificar clientes leales
- Identificar clientes con probabilidad de cambiar de categoría
- Encontrar correlaciones entre indicadores financieros
- Identificar reglas de marcados de valores

Seguros y Salud

- Análisis de procedimientos médicos solicitados en conjunto
- Identificar clientes para nuevos servicios
- Identificar patrones de comportamiento en clientes con riesgo
- Detectar comportamiento fraudulento

Transportes

- Determinar la logística de la distribución
- Analizar patrones de carga

Medicina

- Identificar de terapias médicas satisfactorias para distintas enfermedades.
- Asociación de síntomas y patologías
- Estudio de factores de riesgo/salud
- Segmentación de pacientes para atención inteligente del grupo
- Estudios epidemiológicos
- Análisis de rendimiento de campaña de información, prevención
- Predicción de requerimientos de los centros asistenciales para la asignación optima de recursos

Seguridad

- Detección de fraudes
- Reconocimiento fácil
- Sistemas biométricos
- Acceso a redes no permitidas

Astronomía

- Identificar nuevas estrellas y galaxias

Geología, minería, agricultura y pesca

- Identificación de áreas de uso para distintos cultivos
- Identificación de áreas de uso para pesca
- Identificación de áreas de uso para minería
- Exploración de la minería de bases de datos de imágenes de satélite

Ciencias ambientales

- Identificación de modelos de funcionamiento de ecosistemas naturales y/o artificiales
- Para mejorar la observación, control y gestión ambiental

Ciencias sociales

- Estudio de flujos de la opinión publica
- Planificación de ciudades
- Identificación de barrios con conflictos en función de valores socio demográficos

(RIQUELME, 2006).

Sector industrial

- Optimalización centrales eléctricas
- Control de trenes de laminado
- Optimalización de altos hornos
- Optimalización de la producción de cartón
- Gestión de alarmas de plantas petroquímicas
- Control de calidad en la fabricación de electrodomésticos
- Optimalización de procesos de producción de cemento

Administración publica

- Análisis y control de tráfico de vehículos
- Predicción de demanda de tiempos de trabajo para reparto portal
- Predicción de flujos de turismo

(DAEDALUS, 2006).

5.5.2 Retos y tendencias de la minería de datos.

Existen algunos retos que superar antes de que la minería de datos se convierta en una tecnológica de masas (KARGUPTA, 2004). Aquí se menciona algunos de los retos actualmente planteados.

En cuanto a los aspectos metodológicos sería muy útil la existencia de una API Estándar, de forma que los desarrolladores puedan integrar sin dificultad los resultados de los diversos algoritmos de minería.

Esto podría facilitar también la tarea de automatizar y simplificar todo el proceso integrado en los aspectos como muestreo, limpieza de datos, minería, visualización, etc. En este mismo sentido sería deseable que los productos de minería de datos estuvieran orientados al programador para fomentar su uso y ampliación. Seria asimismo necesario unificar la teoría sobre la materia así se puede observar que los estados del arte son generalizables, no existe un estándar para la validación de resultados y, en general, la investigación se realiza demasiado aislada.

También es necesario mejorar la formación en esta área entre los titulados universitarios, que seria la mejor manera de expandir su uso, y finalmente, sigue siendo un asunto pendiente la integración del conocimiento del dominio en el algoritmo, y viceversa, es decir, mejorar la interpretación y facilidad de uso del modelo hallado.

La escalabilidad de la minería de datos hacia grandes volúmenes de datos es y será siempre una de las tendencias futuras, ya que el volumen de información que se ha de tratar crecer de manera exponencial, con lo que los avances en esta área, quedan siempre superados por las necesidades crecientes.

Datos con miles de atributos es ya algo habitual, pero es probable que las técnicas no estén preparadas aun para centenares de miles o incluso millones de características. Dentro de esta línea también se localiza la minería de data stream de muy alta velocidad con posibles cambios de estructura,

Dimensión o modelo de generación dinámico durante la fase de entrenamiento. Esto obliga a tener un modelo de conocimiento en todo momento.

Habría que hacer una simulación, integración en la toma de decisiones y minería de datos. Básicamente se trata de utilizar las salidas de unos modelos como entradas de otros y maximizar el beneficio del conjunto de modelo. Además, pueden añadirse al modelo global restricciones de valores máximos o mínimos (saturación), etc.

Las técnicas tradicionales de combinación de modelos no pueden aplicarse directamente (ESTRUCH, 2044). Las técnicas de simulación en minería de datos, mas relacionadas con el problema de una maximización global no han recibido la atención suficiente desde el área de la minería de datos.

La obtención de modelos que globalmente se comporten bien y que se mantengan dentro de unas restricciones, requieren no solo de matrices de costes y de técnicas como el análisis, sino de otros tipos de métricas y técnicas para el aprendizaje y la evaluación.

La predicción local más idónea para un problema puede implicar la elección de una menos idónea para otro, mientras puede existir una decisión global mejor. Si bien este tipo de decisiones globales han sido estudiadas por la teoría de la decisión (BENHAIM, 2001), por el área de planificación en inteligencia artificial (RUSELL, 2002), esta interrelación entre modelo predoctores, su aprendizaje y problemas de optimización y planificación no ha sido estudiada a fondo.

La aplicación de la minería para datos con una estructura compleja, ya que en numerosas ocasiones los datos procedentes de aplicaciones del mundo real no tiene una representación directa en forma de una única tabla, sino que deben ser representadas mediante estructuras jerárquicas (árboles), interrelacionadas (grafos), conjuntos, etc. Por lo tanto, el reto que se lanza a la comunidad científica que investiga el aprendizaje automático y minería de datos, es el adaptar o proponer nuevas técnicas que permitan trabajar directamente con este tipo de representación.

En este campo también entraría la minería de datos distribuida, donde los datos no se encuentran en una única localización sino como es cada vez más habitual en una red de computadores.

Un caso particular seria la minería de datos multimedia, para datos que integran voz, imágenes, texto, video, y que, debido a la complejidad de los datos, el volumen y el gran abanico de aplicación posible constituye un reto en la actualidad (YANG, 2005).
Otros temas que se están abordando y donde se debe profundizar son: la comprensibilidad de los patrones extraídos; potenciar las aplicaciones en campos nuevos como privacidad, anti-terrorismo, crisis energética, medioambiente, bioinformática, asegurar la privacidad e integralidad de los datos sensitivos al coste, no solo en el error al asignar una clase sino en la obtención de los atributos; datos en secuencia y series temporales cada vez más utilizadas; etc.

Podemos concluir señalado que la minería de datos se considera todavía un nicho y un mercado emergente. Una de las razones es que la mayoría de los paquetes de minería de datos están dirigidos a expertos, y esta cuestión no facilita su uso por los usuarios. Se piensa que en los próximos años habrá más desarrolladores de aplicaciones comerciales de gestión que sean capaces de integrar en estos, módulos de minería de datos. Con ello se conseguida extender y generalizar su uso a usuarios de los mas diversos campos de la actividad humana.

5.6 DOCUMENTACION DE DOS APLICACIONES

5.6.1 Customer Relationship Management (CRM)

El CRM, es un modelo de negocios cuya estrategia esta destinada a lograr identificar y administrar las relaciones en aquellas cuentas más valiosas para una empresa, trabajando diferentemente en cada una de ellas de forma tal de poder mejorar la efectividad de sus clientes. En otros términos, el CRM se enfoca en que las organizaciones deben ser más efectivas en su interacción con los clientes.

El concepto CRM, se basa en el uso de las más avanzadas herramientas de la tecnología de la información, porque integra la planificación estratégica, las técnicas y herramientas de mercados mas avanzadas con el fin de construir relaciones internas y externas que incrementan los márgenes de rentabilidad de cada cliente y de esta manera, valorar la relación que se establece con ese cliente en el largo plazo para incrementar la rentabilidad de su compañía.

El CRM representa una novedosa tendencia que permite proporcionar la información a escoger y manejar de manera individual a los clientes, con el fin de optimizar su valor para la compañía a largo plazo.

La filosofía del CRM, es un modelo de gerencia que pone al cliente en el centro de los procesos y prácticas de la compañía.

El objetivo primario del CRM, debe ser obtener mayores ingresos y no recortar costos, por lo tanto, se puede afirmar que las soluciones de CRM mejoran los esfuerzos de ventas y de mercadeo, y le permite a las organizaciones proporcionar un mejor servicio a los clientes, se ganan nuevos clientes, se retienen los existentes y compran en mayor cantidad. Los usuarios finales se benefician al recibir un mejor servicio y obtienen los productos y servicios que quieren, cuando los quieren.

CRM supone una orientación estratégica de la empresa hacia el cliente. No se trata de implementar una determinada tecnología, ni de crear un departamento para ello, sino que se debe implicar a cada uno de los trabajadores de la compañía con la independencia el papel que desempeña en ella. Con esta orientación totalmente centrada en el cliente es necesario que el CRM se apoye sobre tres pilares fundamentales, tecnológicos; data Warehouse y data Ming.

La tecnología CRM tiene que ser capaz de recoger toda la información surgida de la relación con el cliente con independencia del canal por donde de ha producido: fax, e-mail, fuerza de venta, Internet, teléfono y analizar para así conocer sus necesidades y poder satisfacerlas.

Con respecto al data Warehouse necesita para una óptima explotación de los procesos la modelización de la información. De esta forma, se establece relaciones causales entre los datos con un objetivo de negocios predeterminados.

El data mining es una herramienta tecnológica eficaz basada en la aplicación de técnicas analíticas y estadísticas a una población de datos registrada en la data Warehouse.
La finalidad del data mining es obtener patrones de comportamiento entre determinados conceptos de información de los clientes. Entre otros se puede prever la demanda,
Analizar la cartera de productos, hacer una simulación de precios, simulación de campañas o investigar y segmentar mercados. (MORA, 2044).

Por consiguiente, le ayuda a seleccionar los preceptos en los que su organización debe concentrar los esfuerzos, a ofrecer a sus clientes los productos adicionales adecuados y a identificar a los buenos clientes que pueden estar a punto de dejarlo.

Esto trae como resultado una mejora en los ingresos, debido a que se mejora la capacidad de responder a cada contacto individual de la mejor manera posible, y a que se reduce los costos ya que se asignan los recursos de manera eficiente.

Las aplicaciones de CRM que usan la minería de datos son llamadas CRM analítico.
El data mining es usado frecuentemente para asignar una calificación aun cliente o prospecto particular que indique la probabilidad de que un individuo se comporte de la manera que usted espera. Por ejemplo, la calificación podría medir la propensión a responder a una oferta particular, o la propensión a cambiarse a un producto/servicio ofrecido por su competencia. También es usado con frecuencia para identificar un grupo de características (llamado perfil) que segmente a los clientes en grupos con comportamientos similares, por ejemplo grupos que compren un producto en particular.

- El CRM (Customer Relationship Management) este modo simplemente significa administrar todas las interacciones con los clientes. En la práctica, esto requiere usar la información acerca de sus clientes y prospectos para interactuar mas rentablemente con ello en todas las etapas de su relación con ellos.

Data mining puede mejorar su rentabilidad en cada una de estas etapas cuando usted lo integra con los sistemas de CRM operacional o lo implementa como aplicaciones independientes (EDELSTEIN, 2005).

Esta aplicación es diseñada por Microsoft y esta disponible en las siguientes ediciones:

- Microsoft CRM 3.0 Professional Edition
- Microsoft CRM 3.0 Small Bussines Edition

Las características comunes de ambas ediciones de Microsoft CRM son:

- Administración de cuentas y contactos
- Organización jerárquica de las cuentas
- Calendario
- Notas y datos adjuntos
- Correo electrónico directo
- Administración de actividades y tareas
- Actividades configurables
- Buscar y búsqueda avanzada
- Combinar
- Mejorar de impresión (comparado con Microsoft CRM v1.2)
- Informes
- Informes con parámetros
- Administración de oportunidades
- Administración de clientes potenciales
- Combinación de correspondencia
- Administrar zonas
- Cliente de Microsoft CRM para Outlook (dos versiones, una con conexión y otra sin conexión)
- Cuotas
- Ofertas, pedidos y facturas para servicios
- Flujo de trabajo
- Catalogo de productos
- Seguimiento de competidores

- Documentación de ventas
- Campañas e informes
- Contratos
- Enrutamiento de clientes potenciales
- Administración de casos
- Administración de Knowledge base
- Mejoras en Knowledge base
- Almacenamiento en cola de actividades y casos
- Administración de correo electrónico, incluidas respuestas automáticas
- Automatización de marketing
- Mejoras en administración de servicios
- Servicio de citas
- Calendario de trabajo
- Administración de trabajo
- Notificación sobre la programación
- Enrutamiento de casos

(MICROSOFT, 2005)

5.6.2 WebMining log Sessionizator

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, como es la navegación de un cliente antes de hacer una compra en línea.

El web Ming log sessionizator, es una herramienta de procesamiento y análisis, que permite la generación de reglas para comprender el comportamiento de los visitantes de su sitio web.

La siguiente figura muestra una visita general de la utilización de WSL XP dentro del esquema de funcionamiento de un sitio web:

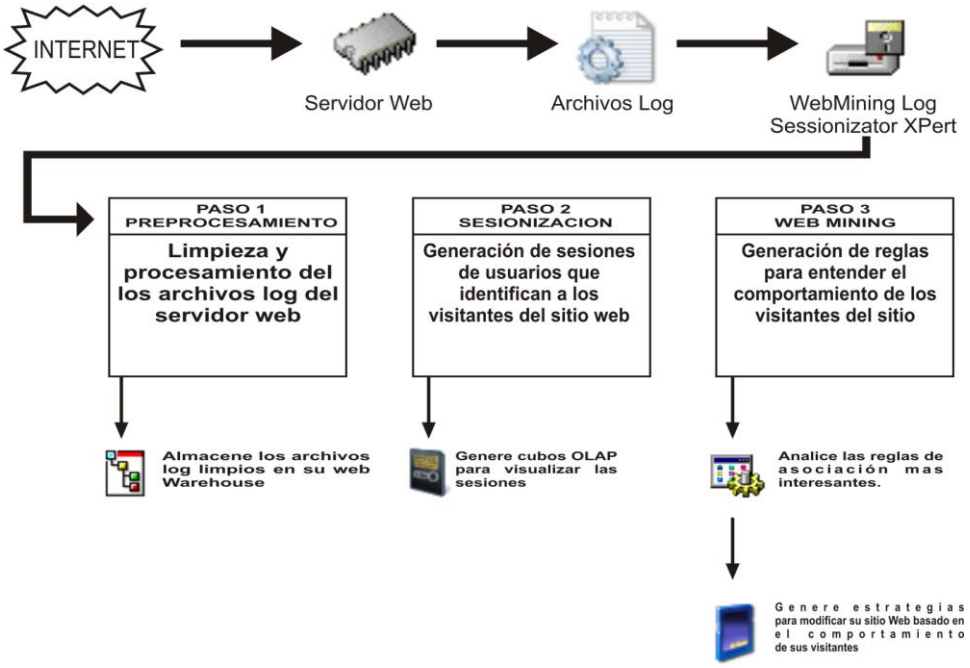

Figura 5 Proceso del web mining log senssionizator.

En solo 3 pasos se puede transformar los datos en bruto de su sitio web en información útil para web mining y analizar la actividad de los usuarios de su sitio web generado reglas dinámicamente. El siguiente diagrama describe los pasos que se deben seguir pata utilizar WSL XP:

- Paso 1.Pre procesamiento Procesamiento de los archivos log del servidor web.
- Paso 2.Sesionizacion Transformación del archivo log en sesiones de usuario.
- Paso 3. Web mining Obtención de reglas que describen el comportamiento de los visitantes de su sitio.

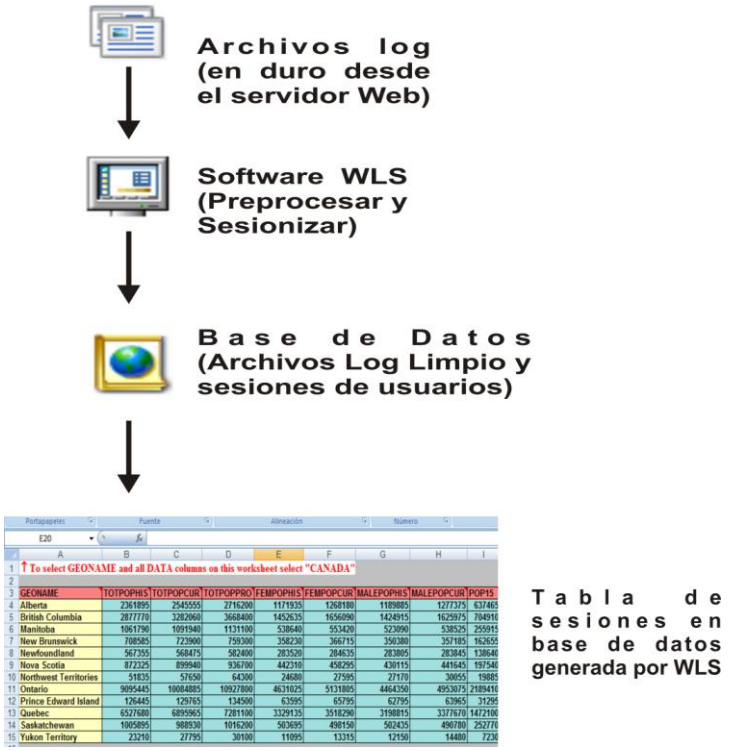

Figura 6. Pasos del web mining log senssionizator

Características

- Mejor rendimiento.
- Pre procesamiento más rápido.
- Sesionizacion más rápida.
- Soporte para variados log files.
- Asistente de proyecto.
- Asistente para reglas de asociación.
- Reporte de reglas estándar y extendidas.
- Scrip a partir de reglas.

(WEBMINING, 2006).

VI. CONCLUSIONES

La mayoría de las compañías tienen una gran cantidad de datos almacenados en sus ordenadores. Estos datos contienen una información que puede ser de gran utilidad para los resultados de la empresa. La gran abundancia de estos o su deficiente estructura puede hacer muy difícil extraer esta información útil.

El data mining era un tema poco conocido hace tres décadas, pero ahora ha tomado una importancia impresionante dentro de las tecnologías de información.

El data mining como se pudo describir en este trabajó es un proceso compuesto por un conjunto de técnicas estadísticas y de inteligencia artificial con el propósito de descubrir patrones ocultos, relaciones interesantes, pronosticar comportamientos, y generar nuevos conocimientos para ayudar a las empresas en la toma de decisiones.

- El objetivo del data mining es la extracción de forma automática de información relevante, útil y no evidente contenida en dichos datos.

Un sistema data mining analiza las fuentes de almacenamiento de información de una organización para encontrar situaciones relevantes que apoyen a la estrategia de una empresa. Debido a que el data warehouse es una tecnología que ha revolucionado la administración, almacenamiento y análisis de información de una empresa, esta tecnología es altamente recomendable para que sea la fuente de información de un data mining.

Como nos pudimos dar cuenta el data warehouse es un sistema de almacenamiento diseñado para el análisis de datos de todos los sistemas de información de las organizaciones, este llegar a ser confundido con un sistema de información de las organizaciones, este llega a ser confundido con un sistema de base de datos lo cual es incorrecto, ya que el data warehouse es diseñado de acuerdo a la estructura del negocio y tiene un enfoque analítico, mientras que un sistema de bases de datos es meramente transnacional.

Se puede observar que un data warehouse puede ser diseñado de dos formas, ya sea de forma central para todos los departamentos de la organización, o en pequeñas partes enfocadas a cada área de la empresa los cuales son llamados data marts.

Otra tecnología que no podemos olvidar ya que es un apoyo importante en el proceso de data mining, es la denominada OLAP, ya que por medio de esta herramienta el usuario puede hacer pequeñas consultas en los data warehouse.

En realidad data mining y KDD son confundidos a menudo y podemos concluir también que el data mining es una herramienta del KDD ya que este es en si el descubrimiento de información en las bases de datos.

En este trabajando se dio a comprender que el data mining se puede aplicar a cualquier tipo de empresa, sin embargo se pudo apreciar que en donde se está desarrollando mas es en el área de marketing, bancos, aseguradoras, transporte y medicina.

En los últimos años el crecimiento del uso del data mining, se debe a que esta herramienta es muy usada en sistemas de CRM y administración del conocimiento, por lo que mientras más empresas mexicanas adopten esta tecnología el data mining será más empleado sector mexicano.

El data mining ha sido implementado por muy pocas empresas en México debido a que es una herramienta relativa nueva, muy costosa en cuestión de software y hardware, las empresas no están todavía preparadas para el uso de herramientas analíticas y porque hay pocos especialistas en el país, sin embargo comparado la inversión que representa la aplicación de data mining, con los resultados obtenidos con el uso de ella es prácticamente justificable dicha inversión.

Las pocas empresas mexicanas que han implementado esta técnica se basa en tres razones principales las cuales son que necesitan general una ventaja competitiva ante sus rivales, a la falta de aprovechamiento de la información que presentan las empresas en el país y particularmente para conocer mejor a sus clientes.

Sin duda el data mining es una de las aplicaciones más interesantes y con una mayor futuro en el área de las tecnologías de la información, ya que en un tiempo no muy lejano va a volverse tan común y fácil de usar como el e-mail y sin duda será la clave del éxito en el área de los negocios.

VII BIBLIOGRAFÍA

DÍAZ Arévalo José Luis; PEREZ García Rafael, 2002, "Estado del arte en la utilización de técnicas avanzadas para la búsqueda de información no trivial a partir de datos en los sistemas de abastecimiento de agua portable", Universidad Politécnica de Valencia, España.

MOLINA Feliz Luis Carlos, 2002ª, "Data mining: Torturando los datos hasta que confiesen", Universidad abierta de Cataluña, España.

CABENA P.; HADJINIAN P, "Discovering data mining from concept to implementation book", Prentice hall. NJ.

BIGUS Josep P., 1996 "Data mining with neural networks", Mc Graw Hill.

PEACOCK P.R., 1998, "Data mining in marketing Part 1" , Marketing Management.

FAYYAD U.N., "Data mining and knoweledge discovery"

MOLINA Félix Luis Carlos; RIBEIRO S., 2001 "Descubrimiento para el mejoramiento bovino usando técnicas de data mining"., Congreso Catalán de ia, Barcelona.

WITTEN C.; FRANK H.; CLARK P.; BOSSWELL R., 2000, "Data mining practical machine learning tool and technique with java implementations" Morgan Kaufmann Publishers.

EDELSTEIN Herbert, 1999 "Introduction to data mining an knowledge discovery", third edition, Two Crows Cosporation, USA.

GOEBEL M., 1999 "A survery of data mining and knowledge discovery software tools".

HERNÁNDEZ Orallo José, RAMIREZ Quintana and FERRI C., 2004 "Introducción a la minería de datos", Prentice addison Wesley.

ANSWERMATH, 2005, "Algunas herramientas utilizadas en la minería de datos" on-line, URL:http://ww.answermath.com/data-mining/mineria-de-datos-6-herramientas.htm

GARRIDO Lluis; LA TORRE José Ignacio, 2001 "Aplicación empresariales de data mining ", Universidad de Barcelona, España.

AGGARWAL & YU´S, 1998 "Data mining techniques for associations, clustering classification", Lecture notes in computer science.

GAMBERG D., SMUC tomislav and MARI Ivan., 2001 "Data mining server", Rudjer boskovic institute.

BERRY Michael & GORDON Linoff, 2000 "Data mining techniques for marketing sales, and customer support", wiley, EUA.

KALAKOTA Ravi & ROBINSON Marcia, 2000, "e-Busines 2.0 Roadmap for success", Addison Wesley, EUA.

ADRIAANS Piet & ZANTINGE Dolf, 1997, "Datamining", Addison wekey, Inglaterra.

CASTAÑEDA Corvera oraldo, 2001 "Identificación de los elementos criticos de éxito en la implementación de la fuente de información: DW", ITESM Monterrey.

DATABASE INC, 2000 "Datawarehousing An executive´s perspectiva",on line, URL:http://www.dspace.com/whatman.htm,2000.

MEJÍA Cosio alberto, 1996, "Datawarehousing. Idea economica", Info latina, México

MICROSOF Corporation, 1996 , "Microsoft Authenticode tecnology", on line, URL:http://www.microsoft.com/security/tech/misf8htm.

INMON H. William H. Anmon, HACKATHOTN Richard D.,1994, "Using the data warehouse", John Wiley & sons.

INOM H William, 1996 "Building the data warehouse ", john Willey, new cork USA:

SANCHEZ Jorge y CRIADO Alberto, 2002, "Dimensiones en un data warehouse", on line , URL:http.//www.infodata.es.

GUPTA R. Vivek, 1997, "An introduction to datawarehousing", system, services corporation papers.

Sas institute, 2001, "Rapid data warehousing with the SAS System", URL:http:/www.sas.com.

FUTUREANALYTICS, 1988, "Identifyning situations where it is advantageous to use Data mining", on line, URL:http://www.futureanalytics.com/papers.htm.

INEI, 1997 "manual para construer un datawarehouse". Instituto Nacional de estadística e informática, lima, Perú.

JOHNSON Amy,1999, "Data Warehousin", Computerworl.

MARCH Bernardo de Quiros y LUNA, 2001, "Tecnologias OLAP en la gestión de servicios de salud: su aplicación en el estudio del gasto en medicamentos", on line URL:http://www.sis.org.ar/tlibres/B/b_13.pdf.argentina.

FAYAD Usama y PIATESKY Shapiro Gregory, 2000, "Advances in Knowledge discovery and data mining", The MIT Press, EUA.

BRACHMAN R.J. KHABAZA T., KLOESGEN W., PIATETSKY-SHAPIRO G., SIMOUDISE E., 1996 ".mining business databases", Communications of the ACM.

FRAGOSO Pérez Carreño Elionora, 2001 "Lineamientos necesarios para la implantación de data mining en las empresas en regiomantanas", ITESM, Monterrey.

RODAS J., 2001 "Un ejercicio de análisis utilizando rouge sets en un dominio de educación superior mediante el proceso KDD", Universidad politécnica de Cataluña. España.

GUZMAN Arenas Adolfo MARTINEZ Luna Gilberto, 1999 "Minería de datos con búsqueda de patrones de comportamiento", Boletín de política informática, INEGI, México.

RIQUELME J.C., RUIZ R. y GILBERT Karina., 2006 "Minería de datos: Conceptos y tendencias", Revista Iberoamericana de Inteligencia Artificial, Sevilla, España.

DAEDALUS, 2006, "Data mining". DAEDALUS-Data, decisions and Languaje.S.A. on line, URL:http://www.daedalus.es/areasmd-e.php.españa.

KARGUPTA H. Joshi, SIVAKUMAR K., YECHA Y., "Data mining next generation challenges and future directions", MIT/AAAI, 2004

ESTRUCH V., HERNANDEZ Orallo J., RAMIREZ J., 2004, "Bagginng decisión multitress", multiple classifier systems.

BEN-HAIM, 2001 "Decision Theory", Academic Press.

RUSELL J . and NORVIG P., 2002 "Artificial intelligence: A modern Approach", prentice Hall.

YANG Q and Wu X., 2005 "Challengin problems in data mining", ICDM. On line URL:http://www,cs.ust.hk/gyang.

MORA Vanegas Carlos, 2004, "La gerencia de marcados y el CRM", UC.

EDELSTEIN Herb, 2005 "Construya relaciones rentables con sus clientes", Two crows corporation. USA.