UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

FACULTAD DE INFORMÁTICA

DATA MINING

TESINA

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN INFORMÁTICA

PRESENTA:

AMPARO ARMIDA LÓPEZ MARTÍNEZ

MAESTRO ASESOR:

ING. JUAN GABRIEL FRANCO DELGADO

SANTIAGO DE QUERÉTARO, QRO. , SEPTIEMBRE DE 2000.

CARTA DE ACEPTACIÓN

Por este medio, se otorga constancia de aceptación de tesina para obtener el título de Licenciado en Informática, que presenta la pasante **AMPARO ARMIDA LÓPEZ MARTÍNEZ** con el tema denominado *Data Mining.*

Este trabajo fue desarrollado como una investigación derivada del curso de titulación **"SISTEMA OPERATIVO UNIX I" – Nivel Introductorio –,** dando cumplimiento a uno de los requisitos contemplados en el artículo 34 del reglamento de titulación vigente, en lo referente a la opción de titulación por realización y aprobación de cursos de actualización.

Se extiende la presente para los fines legales a que haya lugar y para su inclusión en todos los ejemplares impresos de la tesina, a los veintitrés días del mes de febrero del dos mil uno.

ATENTAMENTE

ING. JUAN GABRIEL FRANCO DELGADO
PROFR. CURSO DE TITULACIÓN

INTRODUCCIÓN

El objetivo fundamental de este trabajo, es mostrar al Data Mining como una nueva tecnología, con gran potencial para ayudar a las compañías a concentrarse en la información más importante de sus Bases de Información (Data Warehouse). Las herramientas de Data Mining predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y conducidas por un conocimiento acabado de la información.

Las herramientas de Data Mining pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas y a los cuales los usuarios de esta información casi no están dispuestos a aceptar. Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque se encuentra fuera de sus expectativas.

Las técnicas de Data Mining pueden ser implementadas rápidamente en plataformas ya existentes de software y hardware. Una vez que las herramientas de Data Mining son implementadas en computadoras cliente servidor de alto rendimiento o de procesamiento paralelo, pueden analizar bases de datos masivas y presentar los resultados en formas de tablas, con gráficos, reportes, texto e hipertexto.

1. DEFINICIONES

- " Se puede interpretar como el proceso, máximamente optimizado, intermedio entre la información y la toma de decisiones asociada a la misma. La aplicación ideal del DM se llevaría a cabo sobre las bases de datos corporativas, que pueden ser un Data Warehouse. " (Borrajo Díaz)

- ***" Es la extracción de información oculta y predecible de grandes bases de datos***. Es una poderosa tecnología nueva con gran potencial para ayudar a las compañías a concentrarse en la información más importante de sus Bases de Información (Data Warehouse). Las herramientas de Data Mining predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y conducidas por un conocimiento acabado de la información. "

- " Es el proceso de extracción de información significativa de grandes bases de datos, información que revela inteligencia del negocio, a través de factores ocultos, tendencias y correlaciones para permitir al usuario realizar predicciones que resuelven problemas del negocio proporcionando una ventaja competitiva. "

2. FUNDAMENTOS

La necesidad paralela de motores computacionales mejorados puede ahora alcanzarse de forma más efectiva con tecnología de computadoras con multiprocesamiento paralelo. Los algoritmos de Data Mining utilizan técnicas que han existido por lo menos desde hace 10 años, pero que sólo han sido implementadas recientemente como herramientas maduras, confiables, entendibles que consistentemente son más funcionales que métodos estadísticos clásicos.

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

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

3. EL ALCANCE

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

Predicción automatizada de tendencias y comportamientos.

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

Descubrimiento automatizado de modelos previamente desconocidos.

Las herramientas de Data Mining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso. Otros problemas de descubrimiento de modelos incluye detectar transacciones fraudulentas de tarjetas de créditos e identificar *datos anormales* que pueden representar errores de tecleado en la carga de datos. (Presser 2000)

4. LA INFORMACIÓN COMO UN FACTOR DE PRODUCCIÓN

Muchas organizaciones internacionales producen mas información en una semana que alguna persona podría leer en su vida. Esta es cada vez más alarmante en redes de área amplia como la Internet. Cada día, cientos de megabytes de datos son distribuidos alrededor del mundo, pero no es posible monitorear esta rápido desarrollo, el crecimiento es exponencial.

Estamos enfrentándonos a la nueva paradoja del aumento de datos, donde mas datos significan menos información. En el futuro, solo la habilidad de leer e interpretar no será suficiente para sobrevivir como profesional, un científico o una organización comercial. La Producción mecánica y reproducción de datos nos obliga a adaptar

nuestras estrategias y desarrollar métodos mecánicos para filtrar, seleccionar e interpretar datos.

Las organizaciones que son excelentes haciendo esto, tendrán una mejor oportunidad de sobrevivir, y por esto, su misma información llega a ser un factor de producción de importancia. Esta tendencia es quizás la más obvia en la bolsa, donde esto no es únicamente la disponibilidad de datos que es vital pero también la habilidad de interpretar los datos, y el acto en las bases de estas interpretaciones. La operación de la bolsa tiene, un grado grande, llega a ser un juego de computadoras contra computadoras, con intervenciones humanas solo en el meta-nivel. (Presser, 2000)

Cuando los analistas empresariales utilizan el Data Warehouse para determinar lo que están haciendo sus clientes, una importante pregunta cruza por su mente. ¿Porqué lo hacen? Comprender la conducta de los clientes o el comportamiento empresarial es fundamental para mejorar el balance de la empresa y tener clientes complacidos. Los administradores y analistas empresariales buscan respuestas para lograr objetivos como los siguientes:

- Localizar y llegar a mejores clientes.
- Descubrir nociones empresariales vitales que ayuden a controlar la participación en el mercado y elevar las utilidades.

- Comprender la relación total con cada cliente para desarrollar las estrategias de precios adecuadas y el mayoreo de productos correcto, con base no sólo en la intuición, sino en el uso real del producto y la experiencia del cliente.
- Discernir un valor de por vida para el cliente.
- Reducir los gastos promocionales e incrementar al mismo tiempo la efectividad neta de las promociones en general.

5. INGREDIENTES DEL DATA MINING

Para lograr éstos objetivos, el Data Warehouse proporciona al gerente empresarial dos ingredientes esenciales. El primero es una gran cantidad de datos sobre sus clientes, así como la historia entre el cliente y la organización. El segundo, mucho más importante, es el carácter único de estos datos - ninguno de los competidores los posee. La solución del Data Warehouse debe incorporar la minería de datos a su plataforma de soporte de decisiones.

El Data Mining es un arma esencial en el arsenal del soporte de decisiones del analista. Auxilia a los usuarios empresariales en el procesamiento de vastas reservas de datos para descubrir relaciones insospechadas.

Los analistas empresariales tienen un rango de necesidades. La primera necesidad es comprender que está sucediendo en el negocio. La siguiente es porqué está sucediendo, ¿Cuál es el comportamiento de clientes y mercados?. La última necesidad es ¿Qué puede hacerse? ¿Cuáles acciones se pueden tomar? El valor de un análisis para los gerentes es más alto cuando genera una recomendación factible. Comprender el comportamiento y los pronósticos de clientes y mercados, y lo que puede hacerse, son retos para las técnicas tradicionales de análisis. Las consultas, reportes y análisis multidimensional tradicionales se concentran en lo que está sucediendo y, en menor medida, en el porqué. El Data Mining se concentra en llenar la necesidad de descubrir el porqué, para luego predecir y pronosticar las posibles acciones con cierto factor de confianza para cada predicción.

Las herramientas de Data Mining son un componente importante del sub-bloque de análisis y reportes del bloque de acceso. Este es utilizado para tener una interfaz con el Data Warehouse y con el mercado de datos. Muchas de las herramientas del Data Warehouse también emplean el componente de depósito local del bloque de acceso y recuperación, a fin de almacenar los datos en estructuras de datos de propietario para análisis subsecuentes y presentaciones de los resultados. La mayoría de las herramientas de Data Mining puede con facilidad saltarse el Data Warehouse o el mercado de datos y acceder de manera directa la fuente de los datos. Tradicionalmente, las herramientas de Data Mining acceden los datos de la fuente, sin embargo, los datos del Data Warehouse o del mercado de datos están refinados,

integrados y estandarizados. La estandarización eliminó aspectos como las convenciones de nombres múltiples, las estructuras ocultas de codificación y los campos faltantes. Los datos operacionales en la fuente son por lo general inconsistentes y están dispersos en muchas aplicaciones. Además, se requieren datos históricos para descubrir patrones temporales de interés.

El Data Mining difiere en varias formas del procesamiento informático y analítico. En el siguiente cuadro se observan las principales diferencias:

	El procesamiento informático/analítico en contra de el Data Mining	**Minería de datos**
Enfoque	Datos de resumen	Datos de transacción o de detalle
Dimensiones	Limitadas	Muchas
Cantidad de atributos	Total de decenas	Cientos para cada dimensión
Tamaño del conjunto de datos	De reducido a mediano para cada dimensión	Millones para cada dimensión
Enfoque del análisis	¿Qué está sucediendo en el negocio?	¿Porqué está sucediendo? Acciones de predicción y pronóstico
Técnica de análisis	Rebanar y picar	Descubrir automáticamente
Proceso de análisis	Analista empresarial iniciado y controlado	Datos y sistema iniciado Orientación mínima al analista empresarial
Factor de confianza	Derivado por el analista empresarial	Derivado de los datos
Estado de la tecnología	Desarrollada	Desarrollada en análisis estadístico Incipiente en descubrimiento de conocimientos

Los datos en el Data Warehouse deben estar al nivel de detalle correcto. Debido a la naturaleza incipiente de la tecnología de minería de datos, es necesaria - en especial al principio - una estrecha cooperación entre los analistas empresariales y los profesionales en tecnología de la información.

Para formar la mezcla correcta de actividades de minería de datos, son cruciales tres ingredientes: usuarios, aplicaciones empresariales y tecnología y herramientas.

6. USUARIOS DEL DATA MINING

Los usuarios claves del Data Mining son los analista empresariales, los peritos en estadística y los profesionales en tecnología de la información que auxilian a los usuarios empresariales. Quienes obtienen beneficios de los resultados de minería de datos son los gerentes empresariales y los ejecutivos, que desean entender los factores de éxito del negocio con base en datos completos del cliente, y utilizar luego este conocimiento ara afinar las estrategias de producción, precios y comercialización; mejorar el nivel de éxito de las estrategias e impulsar el balance.

Hasta la fecha, las empresas han dependido del procesamiento informático y analítico para medir y comprender la estabilidad del negocio. El procesamiento informático - consultas y reportes - es más sencillo de usar, pero requiere de una estrecha dirección

del analista. El procesamiento analítico (OLAP) requiere de menos dirección del analista, aunque los datos deben estar organizados en una forma especial (base de datos multidimensional), o accederse bien de manera especial (visión multidimensional). En ocasiones se utiliza una combinación de técnicas de consulta y OLAP para comprender el comportamiento del cliente o para construir perfiles de segmentos de mercado; pero el proceso de aplicar estas técnicas es conducido esencialmente por el analista empresarial. En estos casos, este proceso también se conoce como minería de datos. Se ha definido a el Data Mining como la modalidad de descubrimiento del soporte de decisiones, la cual es conducida por los datos y no por el analista empresarial.

Conducida por el Analista	Auxiliada por el Analista		Conducida por los Datos
Procesamiento Informático Consultas Reportes	Procesamiento Analítico OLAP MDDBMS OLAP Relacional	Análisis estadístico/de datos	Descubrimiento de conocimientos

7. PROBLEMAS QUE RESUELVE EL DATA MINING

- Falta De visión a largo plazo: necesita preguntar "¿qué queremos para nuestros archivos en el futuro?".
- No todos los archivos están actualizados en la fecha: los datos son incorrectos o se han perdido; los archivos cambian mucho en calidad.
- Esfuerzo entre departamentos: algunos departamentos no quieren actualizar sus datos.
- Cooperación pobre del departamento de procesamiento de datos: por ejemplo, "dicen darnos las consultas y encontraremos la información que quiere."
- Restricciones legales y de privacidad: algunos datos no pueden ser usados, por razones de privacidad.
- Los archivos son difíciles de conectar por razones técnicas: hay una discrepancia entre una base de datos jerárquica y una relacional, o modelos de datos que no están actualizados.
- Sincronizando problemas: Los archivos pueden ser compilados centralmente, pero con retraso de seis meses.
- Interpretación de problemas: Las conexiones son encontradas en la base de datos, pero no conocen su significado o que ellas pueden ser usadas. No obstante estas

estructuras pueden algunas veces ser muy valiosas; pueden ser indicaciones de contaminación o guiar a la detección de fraude.

Ninguno de estos problemas son insolubles en una flexible y saludable organización. La experiencia nos dice que cuando son dirigidos en la forma apropiada, una organización puede beneficiarse tremendamente de la introducción del KDD. (Harjinder, 1996)

8. SURGIMIENTO DE APLICACIONES DE DATA MINING

En las aplicaciones empresariales, a la fecha, la tecnología de Data Mining se ha utilizado principalmente en aplicaciones de comercialización, ventas y análisis de crédito; y se ha aplicado con éxito en áreas empresariales con el más alto potencial, tales como la segmentación de clientes y del mercado y el análisis de comportamiento del cliente, en particular en los sectores de menudeo, bancario y financiero. Hasta aquí, la tecnología por lo general era costosa de aplicar y desplegar, pero esta situación está cambiando con rapidez. Hoy en día, tres fuerzas importantes conduce el crecimiento del Data Mining:

- La tecnología de Data Warehouse para proporcionar un gran banco de datos bien organizados e históricos.
- Hardware en paralelo, productos de base de datos y herramientas a precios razonables.
- Tecnología y herramientas para minería de datos cada vez más desarrolladas.

Se espera que se acelere el uso del Data Mining. La cantidad de aplicaciones del Data Warehouse crece con rapidez y los precios de hardware en paralelo y los productos de apoyo de software disminuyen con rapidez.

9. DATA MINING Y EL DATA WAREHOUSE

9.1 ¿QUÉ ES EL DATA WAREHOUSE Y PORQUÉ SE NECESITA?

Las organizaciones modernas están bajo presión para responder rápidamente a los cambios en el mercado. Es claro que, se necesita un acceso rápido a todo tipo de información antes de que pueda tomar decisiones lógicas. Para tomar las decisiones correctas para su organización, es esencial ser capaz de investigar el pasado e identificar tendencias relevantes. Obviamente, para cualquier rendimiento y análisis de tendencias debe tener acceso a toda la información necesaria, y su información es principalmente almacenada en grandes bases de datos. La forma más fácil de ganar acceso a estos datos y facilitar la toma de decisiones efectiva es montar un Data Warehouse.

En muchas organizaciones encontrará realmente grandes bases de datos en operación para transacciones normales diarias y algunas de las aplicaciones usarán monitoreo de transacciones.

Estos tipos de bases de datos son conocidos como base de datos operacionales; en muchos casos no ha sido designado un almacén histórico de datos o responde a

consultas pero simplemente para realizar todas las aplicaciones para transacciones diarias.

El segundo tipo de base de datos encontradas en organizaciones es el Data Warehouse. Desarrollada para estrategias de soporte a la decisión y es grandemente reforzado para base de datos de carácter operacionales. (Adrians, 1996)

9.2 DEFINICIÓN DE DATA WAREHOUSE.

- Es el proceso de extraer y filtrar datos de las operaciones comunes de la empresa, procedentes de los distintos subsistemas operacionales, para transformarlos, integrarlos, sumarizarlos y almacenarlos en un depósito o repositorio, y acceder a ellos cada que vez que se necesite. Se puede concebir un Data Warehouse como un almacén-factoría de datos o información, que concentra la información de interés para toda la organización y distribuye dicha información por medio de diversas herramientas de consulta y de creación de informes orientadas a la toma de decisiones. Con esta tecnología se convierten los datos operacionales de una organización en una herramienta competitiva, que permite a los usuarios finales examinar los datos de modo más estratégico, realizar análisis y detección de

tendencias, seguimiento de medidas críticas, producir informes con mayor rapidez, un acceso más fácil, más flexible y más intuitivo a la información que se necesite en cada momento.

- Es un "almacenamiento" de información temática (ejemplo: productos, clientes, etc.) orientado a cubrir las necesidades de aplicaciones DSS y EIS, que permite acceder a la información corporativa para la gestión, control y apoyo a la toma de decisiones.

El **fin** del Data Warehouse o Almacén de datos es reunir y consolidar las bases de datos diferentes, que se mantienen en los diferentes departamentos o áreas funcionales de la empresa como subsistemas de información independientes, en una gran base de datos, recogiendo datos muy dispares y, muchas veces infrautilizados, procedentes de fuentes internas repartidas por toda la organización. También recogerá datos o informaciones externas, que rutinariamente se recibe sobre las diferentes entidades u objetos de información, es decir, clientes, proveedores, productos y servicios, canales, estructura organizativa, competencia, mercado, coyuntura económica, etc., en resumen, los derivados de las relaciones de la empresa con su entorno. (Borrajo, 1999)

La **característica básica** de un Data Warehouse es que contiene una vasta cantidad de datos, los cuales significan billones de registros. Más pequeñas aún, Data Warehouses locales son llamados **DATAMARTS.**

Hay algunas reglas específicas que gobiernan la estructura básica de un Data Warehouse, esto es, cada estructura debe ser:

1. **Dependiente del tiempo:** contiene información coleccionada sobre tiempo, lo cual implica que siempre debe haber una conexión entre la información y el almacén en el momento en que sea introducida. Este es uno de los aspectos más importantes de un almacén como esta relaciona al Data Mining, porque la información puede ser originada de acuerdo al período.
2. **No volátil:** Esto es, que los datos en el Data Warehouse nunca es actualizado, pero es usado sólo para consultas. Así cada dato puede ser cargado de otras bases de datos las cuales son base de datos operacionales. Los usuarios finales que quieran actualizar los datos, deben usar las bases de datos operacionales, como la única forma de actualizar, cambiar o borrar. Esto significa que el Data Warehouse será siempre llenado con datos históricos.
3. **Orientado a temas:** Esto es, construido alrededor de todas las aplicaciones existentes de los datos operacionales. No toda la información en la base de datos operacional es útil para un Data Warehouse, desde el Data Warehouse es diseñado

específicamente para soporte a la decisión mientras la base de datos operacional contiene información de uso diario.

4. **Integrado:** Esto refleja la información de la organización. En un ambiente de datos operacional encontrará algunos tipos de información usada en una variedad de aplicaciones y algunas aplicaciones estarán usando diferentes nombres de las mismas entidades. Por lo tanto, en un Data Warehouse esto es esencial para al integrar esta información y hacerla consistente; sólo un nombre debe existir para describir una entidad individual. (Adrians, 1996)

9.3 COMPONENTES DEL DATA WAREHOUSE.

Fuentes de datos: Este componente es el que normalmente está presente originariamente en las organizaciones, y a partir del cual se realiza la captura de datos que se contemplará en el DW. Estas fuentes de datos pueden ser sistemas operacionales corporativos (representan el entorno del que se obtienen la mayor parte de los datos significativos de la operativa diaria de la compañía), sistemas operacionales departamentales, fuentes externas etc.

Herramientas de acceso: Sin las herramientas adecuadas de acceso y análisis el DW se puede convertir en una amalgama de datos sin ninguna utilidad. Es necesario poseer técnicas que capturen los datos importantes de manera rápida y puedan ser analizados desde diferentes puntos de vista. También deben transformar los datos capturados en información útil para el negocio. Actualmente a este tipo de herramientas se las conocen como business intelligence tool (BIT) y están situadas conceptualmente sobre el DW. Cada usuario final debe seleccionar que herramienta se ajusta mejor a sus necesidades y a su DW. Entre ellas podemos citar las Consultas SQL (Structured Query Language), las Herramientas MDA (Multidimensional Analysis) y OLAP (On-Line Analytical Processing), las Herramientas ROLAP

(Relational On Line Analytical Processing) y las herramientas DATA MINIG, de las cuales se trata a continuación.

Repositorio/Metadatos: Los metadatos son básicamente datos acerca de los datos contenidos en el DW. Así, uno de los problemas con el que pueden encontrarse los usuarios de un DW es saber lo que hay en él y cómo pueden acceder a lo que quieren. El repositorio les ayuda a conseguirlo. Es sólo una de las utilidades del repositorio, pero éste tiene muchas funcionalidades: catalogar y describir la información disponible; especificar el propósito de la misma; indicar las relaciones entre los distintos datos; establecer quién es el propietario de la información; relacionar las estructuras técnicas de datos con la información de negocio; establecer las relaciones con los datos operacionales y las reglas de transformación; y limitar la validez de la información. (Borrajo, 1999)

Diseñar un Data Warehouse requiere conocimientos especializados de diseño de datos porque los modelos de datos consisten de datos necesarios para usuarios quienes desean accesarlos con rapidez, y el diseño de datos para el almacén puede ser completamente diferente de la base de datos operacional. Después de crear un modelo de datos corporativo para el Data Warehouse, tiene que diseñar un ambiente de administración de datos específico. Si se tiene un numero de bases de datos

soportando los datos operacionales, tiene que copiar esta información al Data Warehouse. Debe entonces, ser capaz de controlar este ambiente.

Aún cuando tiene que montar un Data Warehouse, debe apreciar que la estructura de una organización y sus procesos de negocios pueden y cambiarán todo el tiempo. Un sistema de soporte a la decisión es también un sistema que puede cambiar constantemente: si los requerimientos de la organización se alteran, entonces su modelo de datos debe cambiar también. Si modifica los atributos de los datos operacionales, puede influenciar el modelo de datos del Data Warehouse y solo con las apropiadas herramientas de administración de datos pueden controlar este ambiente. Una vez que el Data Warehouse es montado, puede tomar una fotografía instantánea de sus datos y ponerlos en servidores locales de base de datos necesarios. Montar un Data Warehouse es el procedimiento mas apropiado para llevar a cabo soporte a la decisión. Esto lleva a la flexibilidad de hacer consultas ad hoc y también decisiones basadas en datos históricos. En un Data Warehouse, un usuario final puede querer hacer uniones de algunas tablas y esto puede dar lugar a tremendas exigencias en el sistema. Por esta razón, el Data Warehouse requiere una máquina de alta velocidad y una amplia variedad de optimización de procesos.

9.4 METADATO

Montando un Data Warehouse, el usuario final y el administrador deben tener acceso a toda la información en las tablas y los atributos. Querrán conocer un número de cosas, como estos:

- Dónde están localizados los datos.
- Qué datos existen.
- Qué tipo de dato ó formato está.
- Cómo éste dato está relacionado con otro dato en otras bases de datos.
- De dónde es el dato y a quién pertenece.

Por éstas razones, otra base de datos que contiene todo esto el llamada Metadato, el cual describe la estructura del contenido de una base de datos. En un ambiente complejo de base de datos, un adecuado Metadato es indispensable, desde que esto determina la estructura de ambos, datos operacionales y el Data Warehouse. El Metadato es usado por usuarios finales para propósitos de consulta, como también por el administrador de datos para estructurar la administración de un sitio de base de datos.

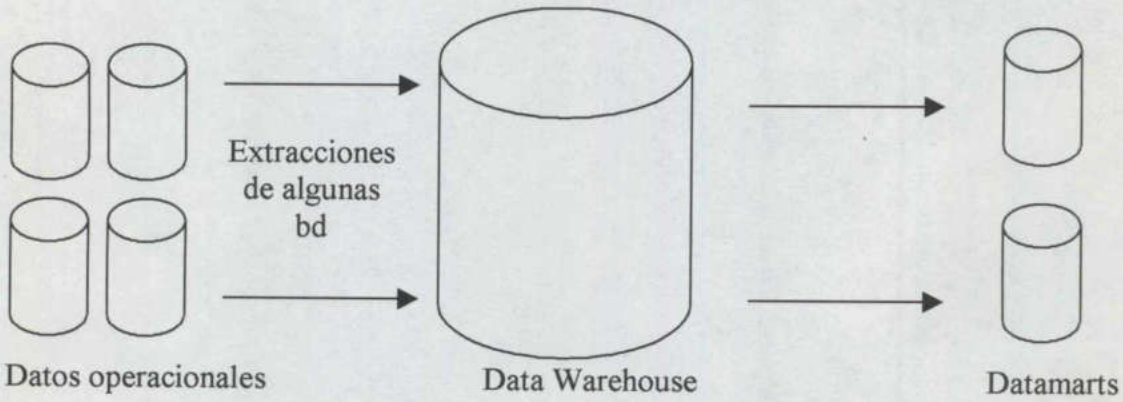
9.5 INTEGRACIÓN CON EL DATA MINING

La aplicación de las técnicas de Data Mining se pueden llevar a cabo en dos formas: por la existencia de un data warehouse o por la extracción del Data Warehouse, de la información de interés para el usuario final y copiando ésta a una computadora específica, posiblemente una máquina multiproceso.

La integración de un Data Mining en un sistema de soporte a la decisión es muy provechoso. La única función de un Data Warehouse es abastecer la información necesaria para tomar decisiones adecuadas. En algunos casos puede usar lenguaje estructurado de consulta estándar (SQL). , pero si se quiere comparar millones de registros y no se conoce exactamente el tipo de información que requiere, o si quiere encontrar datos ocultos, entonces tiene que ir a Data Mining. Algunos tipos de técnicas de Data Mining y cada computadora la usa en formas específicas. Por esta razón, es importante entender las exigencias de los usuarios finales que son capaces de construir un Data Warehouse apropiado para un Data Mining. En algunos casos encontrará que necesita una computadora separada para Data Mining; traer datos operacionales minados es casi imposible porque hay diferentes aplicaciones con diferentes tipos de atributos y diferentes tipos de datos pero no datos históricos. Con un Data Warehouse, este problema no existe, toda la información ha sido transferida

de la base de datos operacional al Data Warehouse; adicionalmente, en algunos casos puede limpiar los datos antes de comenzar a minar datos.

Vemos en la figura, la relación entre datos operacionales, un Data Warehouse y datamarts.



9.6 AMBIENTE CLIENTE SERVIDOR Y EL DATA WAREHOUSE

Desde hace pocos años se ha provisto que es muy díficil construir efectivos sistemas de soporte a la decisión porque las técnicas disponibles no son capaces de soportar al usuario final satisfactoriamente. El usuario final idealmente, debería tener disponibles todo tipo de técnicas como interfaces gráficas al usuario y mecanismos de

ventanas. El ambiente cliente servidor dispersa el software sobre algunas computadoras y crea un ambiente para el usuario final en un solo sistema. La pesada carga de interfaces gráficas de usuarios u otras técnicas visuales pueden ser procesadas en la máquina local y todas las tareas de las bases de datos tomadas por un servidor específico de base de datos. En algunos casos puede comprar base de datos especiales que operen con un tipo específico de hardware.

Todas las técnicas actualmente disponibles en el mercado, representan la mejor opción para construir Data Warehouse bajo ambientes cliente-servidor. También, en el campo del KDD hace sensible al uso de técnicas de Data Mining. Encontrará algunas herramientas cliente servidor que operan como copias, siendo capaz de copiar registros de base de datos operacionales en servidores de base de datos específicos que almacena el Data Warehouse.

Las técnicas de copiado son usadas al cargar la información de la base de datos operacional al Data Warehouse. La técnica que escoja depende en la forma en la cual use su Data Warehouse. Si necesita inmediatamente accesar a la última información, entonces necesita trabajar con las herramientas de copiado mas avanzadas. Si la actualización del Data Warehouse es menos urgente, entonces puede trabajar con actualizaciones por lotes del servidor de base de datos.

El usuario final tampoco puede visualizar el trabajo en la terminal de trabajo local o conectar la pantalla del servidor que tiene acceso al Data Warehouse corriendo en

una o mas servidores de base de datos. Los datos pueden ser extraídos a un sistema de base de datos local, y procesados usando algoritmos de base de datos. Esto es posible con las técnicas de un ambiente cliente-servidor , cada computadora es montada por completo para optimizar la aplicación del usuario final.

Dos técnicas básicas son usadas al construir un Data Warehouse, conocida como método TOP DOWN y BOTTOM UP. En el método TOP DOWN , primero construye un Data Warehouse para la organización completa, el cual será una base de datos enorme donde todos la información del usuario final es almacenada, y de esta selecciona la información necesitada para su departamento o usuario final local. Entonces trabaja en su base de datos con herramientas de Data Mining. En el método Bottom Up, las más pequeñas Data Warehouse locales. Conocidas como datamarts, son usadas por los usuarios finales en niveles locales para sus requerimientos locales específicos, para propósitos de consulta o para análisis estadísticos. Estos usuarios finales no dependen de otros Data Warehouses en la organización. Un Data Warehouse local puede ser construido en un muy corto periódo de tiempo y puede ser manejado en un nivel departamental , cada datamart es completamente optimizado para tareas particulares, tal como Data Mining. Por esta razón, numerosos datamarts son con frecuencia encontrados en una organización. La Data Warehouse corporativa es generada de estas.

Si tiene montado un datamart a ser usado con técnicas de Data Mining, puede optimizar su base de datos local. Primero debe estar seguro que el hardware y los requerimientos de la base de datos que ha establecido son adecuadas para este propósito.

10. MÁQUINAS MULTIPROCESO

Un ambiente de Data Mining tiene específicos requerimientos de hardware. Si un usuario final quiere comparar grandes número de registros dentro de un periódo corto de tiempo, la computadora necesita toda la memoria interna y todo el poder de procesamiento sólo para esta tarea. Cuando trabaja con algoritmos genéticos en particular, es importante de entender las exigencias que son hechas en la computadora: ésta tiene que tomar cada registro, comparar esta con todas los otros millones de registros en la base de datos y encontrando un patrón acertado en la base de datos, recalcular este patrón mientras compara todos los registros constantemente. En casos certeros, el usuario final necesita una respuesta en muy corto tiempo. Además de llevar respuestas satisfactoriamente en por lo menos dos opciones disponibles: define la cuestión enfocándose en un limitado número de registros y atributos sin una base de datos o moverse hacia un sistema de computadora multiprocesamiento.

Todos los registros son almacenados en un disco duro en las que sus máquinas pueden ser usados solo por Data Mining. Hay muchos tipos de máquinas multiprocesamiento y describiremos sólo dos muy importantes:

- Multiprocesamiento simétrico.
- Procesamiento Paralelo

Con el multiprocesamiento simétrico todos los procesadores trabajan sobre una computadora todos son iguales y se comunican por la vía de almacenaje por porción. En máquinas con multiprocesamiento simétrico una comparte el disco duro y la memoria interna. Aunque los procesadores comparten su coordinación interna, este tipo de multiprocesamiento es limitado a un número de procesadores porque la sincronización de los procesadores da lugar a una enorme carga en el sistema; actualmente, aproximadamente veinte procesadores es lo máximo.

La máquina de procesamiento paralelo es una computadora donde cada procesador tiene su propio sistema operativo, su propia memoria, y su propio disco duro. Aunque cada procesador es independiente, es posible la comunicación entre los sistemas. En este tipo de ambiente puede trabajar con miles de procesamientos; sin embargo, con cada parte del software corriendo en un procesador específico, lleva una gran distribución de tiempo para reunir el software en el orden correcto. Cada

computadora es muy útil en muy grandes proyectos de Data Mining, desde los extremadamente poderosos y tiene una muy buena respuesta de tiempo.

Generalmente, los distribuidores de hardware y software producen específicas bases de datos para las máquinas en paralelo, porque la optimización de las consultas en un Data Warehouse y uso de las técnicas así como cada Data Mining requiere un cercana relación entre hardware y software . No todas las bases de datos son capaces de soportar máquinas en paralelo pero las más modernas bases de datos son capaces de trabajar con máquinas paralelas simétricas. Actualmente sólo pocos distribuidores de base datos como IBM con DB/2, Oracle y Tandem, son capaces de operar con computadoras de procesamiento masivo paralelo, pero otras surgirán en su debido momento.

11. TECNOLOGÍAS Y HERRAMIENTAS DEL DATA MINING

Existe una amplia variedad de tecnología para el Data Mining y todavía va a aparecer más en el mercado. Estas herramientas y tecnologías de minería de datos se clasifican en tres grandes categorías:

- Análisis estadístico o de datos.
- Descubrimiento de conocimientos.
- Sistemas de visualización
- Otro como los sistemas de información geográfica, análisis fractal y herramientas de propietario. (Harjinder, 1996)

11.1 DESCUBRIMIENTO DE CONOCIMIENTOS O KDD

En el análisis estadístico y de datos, es esencial que el analista empresarial conozca cuáles son las variables antes de iniciar el análisis. Quizá no se esté satisfecho con el análisis estadístico y se sospecha que los datos ocultan algo. En estas, situaciones, el analista empresarial requiere de la tecnología y las herramientas para el descubrimiento de conocimientos.

El descubrimiento de conocimientos tiene sus raíces en la inteligencia artificial y el aprendizaje con máquinas.

El descubrimiento puede definirse así:

- El descubrimiento de conocimiento es extraer de los datos información implica, no trivial que no se conocía y potencialmente útil.
- El descubrimiento de conocimientos es el proceso de buscar en los datos sin establecer por adelantado una hipótesis o cuestión, e incluso así encontrar información inesperada por adelantado una hipótesis o cuestión, e incluso encontrar información inesperada e interesante de relaciones y patrones entre los elementos de datos o reglas empresariales importantes en todos los datos investigados y analizados.

- El descubrimiento de conocimientos significa descubrir hechos empresariales antes desconocidos los gigabytes de datos del Data Warehouse.

La tecnología para el descubrimiento de conocimientos determina por sí misma las preguntas a formular, y luego continúa formulando preguntas, cavando más a fondo, para desenterrar las pepitas de conocimientos que la empresa busca.

El descubrimiento de conocimientos pretende examinar la vasta cantidad de datos en el Data Warehouse, en busca de patrones recurrentes, detectando tendencias y desenterrando hechos. Los sistemas de descubrimiento de conocimientos intentan descubrir hechos o conocimientos con una mínima o ninguna instrucción u orientación del analista, todo ello en el menor tiempo posible. De modo que, en el descubrimiento de conocimientos, se inspeccionan grandes cantidades de datos del Data Warehouse y se descubren hechos o conocimientos y se presentan al analista empresarial. Después, el analista ejercita su habilidad empresarial y su experiencia en la materia para separar los hechos útiles de los inútiles. El cerebro humano posee los mejores algoritmos para analiza muchas variables al mismo tiempo, pero tiene la amplitud de banda, potencia y paciencia, pero no comprende las diversas variables empresariales. Las herramientas de visualización y exámen de datos que ayudan a explorar y analizar datos extraídos con anterioridad, mejoran el esfuerzo de descubrimiento de conocimientos.

11.2 ESTRUCTURA GENERAL DE LOS SISTEMAS DE DESCUBRIMIENTO DE CONOCIMIENTOS.

Un sistema de descubrimiento de conocimientos consta de una integración de componentes que identifican y extraen patrones y relaciones interesantes y útiles de los datos registrados en el Data Warehouse.

En una herramienta de descubrimiento de conocimientos, los componentes individuales tal vez no están separados. Las principales entradas para el sistema son los datos del Data Warehouse, la guía del analista empresarial y la experiencia en el tema que almacena la base de conocimientos del sistema. Los datos seleccionados del Data Warehouse se procesan en el motor de descubrimiento de conocimientos, en donde se aplican una serie de algoritmos de extracción, para producir prospectos de patrones y relaciones. Estos prospectos se evalúan; algunos se identifican como descubrimientos interesantes y se le presentan al analista empresarial. Algunos de estos descubrimientos pueden agregarse a la base de conocimientos para mejorar las extracciones y evaluaciones posteriores de descubrimientos.

11.3 ADMINISTRADOR DEL SISTEMA DE DESCUBRIMIENTO DE CONOCIMIENTOS

El administrador controla y maneja todo el proceso. Se emplean las entradas del analista empresarial y la información en la base de conocimientos para conducir el proceso de elegir datos, el proceso de seleccionar y usar algoritmos de extracción y el proceso de evaluar el descubrimiento. El administrador del sistema ayuda también a presentar el descubrimiento del analista empresarial y a almacenar los descubrimientos deseados en la base de conocimientos para actividades posteriores. También maneja el nivel de supervisión del motor de descubrimiento de conocimientos. Se requiere mantener un equilibrio entre la autonomía - nivel de guía del analista empresarial y la influencia por lo que almacena la base de conocimientos- y la versatilidad - los tipos de descubrimientos alcanzables y los tipos de temas comprendidos.

11.3.1 BASE DE CONOCIMIENTOS Y ENTRADAS DEL ANALISTA EMPRESARIAL

Los metadatos y los datos del Data Warehouse son entradas para describir las estructuras de datos en Data Warehouse. Un conocimiento adicional de los datos, tal como los campos claves de fecha, que van a enfocar las reglas empresariales para derivar los datos que necesita el análisis y las jerarquías. Las jerarquías de datos, también presentan entradas por parte del analista empresarial a la base de conocimientos. La meta consiste en provocar una búsqueda eficiente de patrones de interés.

El riesgo de influenciar es que puedan ignorarse patrones y relaciones potencialmente útiles. El analista debe establecer un equilibrio. Las herramientas de descubrimiento de conocimientos mejoran almacenando nuevos descubrimientos para conducir el uso siguiente.

11.3.2 INTERFAZ DE BASE DE DATOS DEL DATA WAREHOUSE

Los sistemas de descubrimiento de conocimientos extraen los datos de la base de datos de la Data Warehouse, por medio de las opciones de consulta de la base de datos. Para las bases de datos relacionales se emplea el lenguaje SQL. Los metadatos del Data Warehouse en la base de conocimientos, guían la interfaz de base de datos para corregir la organización de las estructuras de datos y saber cómo se almacenan en el Data Warehouse.

Para efectos de eficiencia y desempeño, la interfaz de base de datos del sistema de descubrimiento de conocimientos debe comunicarse directamente con el Data Warehouse.

11.3.3 SELECCIÓN DE DATOS

Este componente determina cuáles datos se necesitan extraer del Data Warehouse y cuales son sus estructuras de base de datos. La base de conocimientos guía al componente de datos, el componente de selección de datos debe tener la capacidad para extraer y seleccionar la muestra aleatoria correcta. Además, se selecciona e ingresa en el algoritmo el tipo de datos que requiere.

11.3.4 MOTOR DE DESCUBRIMIENTO DE CONOCIMIENTOS

El motor de descubrimiento de conocimientos aplica los algoritmos de extracción de la base de conocimientos a los datos que extrae el componente de selección de datos. La meta aquí consiste en extraer patrones y relaciones entre los elementos de datos. La influencia desarrollada dentro de la base de conocimientos tiene un efecto crítico sobre los descubrimientos extraídos.

Es posible incorporar un amplio rango de algoritmos dentro del sistema de descubrimiento de conocimientos, tales como dependencias de datos, reglas de clasificación, agrupamiento, resúmenes, detección de desviación, inducción y razonamiento confuso.

11.3.5 EVALUACIÓN DE DESCUBRIMIENTOS

Los analistas buscan en los datos patrones interesantes que les ayuden a comprender qué está sucediendo con los clientes. Los productos, el mercado, etc. Un Data Warehouse tiene, en potencia, un sin número de patrones. El componente de evaluación o filtro ayuda al analista empresarial a examinar los patrones para

seleccionar sólo los que sean de interés. Las técnicas que se emplean para analizar patrones interesantes comprenden la significancia estadística, el factor de confianza o nivel de cobertura y el análisis visual.

11.3.6 PRESENTACIÓN DE DESCUBRIMIENTOS

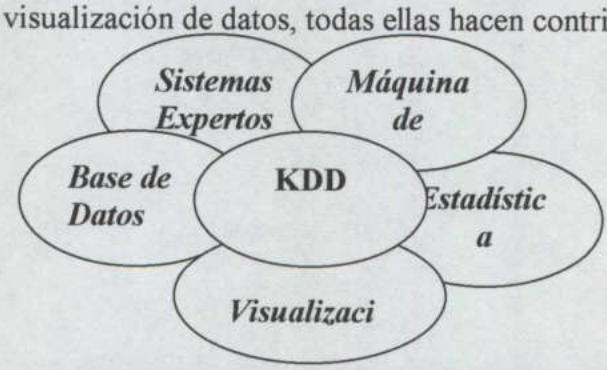
Este componente proporciona dos capacidades: ayudar al analista empresarial a evaluar los descubrimientos, a almacenar los descubrimientos de interés en la base de conocimientos para su futura referencia y su uso, y a comunicar el descubrimiento de gerentes ejecutivos empresariales. Las comunicaciones poco eficaces pueden dañar seriamente el valor del descubrimiento. Aquí lo importante es usar el descubrimiento para comprender el negocio, y convertir luego esta comprensión en recomendaciones factibles. Las técnicas de presentación en un sistema de descubrimiento de conocimientos incluyen la navegación y el exámen visual, el reporte de texto en lenguaje natural y los cuadros y gráficas.

11.4 KDD Y DATA MINING

Existe, en ocasiones la confusión acerca del significado exacto de los términos "Data Mining" y "KDD" puesto que algunos autores los muestran como sinónimos. En la primera conferencia internacional de KDD en Montreal en 1995, fué propuesto que el término KDD es empleado para describir el proceso entero de extracción del conocimiento de datos. En este contexto, conocimiento significa relaciones y patrones entre datos. Fue Adicionalmente propuesto que el término "Data Mining" debería ser usado exclusivamente para el descubrimiento del escenario del proceso KDD.

KDD es la no trivial extracción de lo implícito, previamente desconocido y conocimiento potencialmente útil de los datos.

Así que el conocimiento debe ser nuevo, no obvio, y lo más capaz de usar esto. KDD no es una nueva técnica pero más bien es un multidisciplinario campo de investigación; aprendizaje de máquinas, estadísticas, tecnología de base de datos, sistemas expertos y visualización de datos, todas ellas hacen contribuciones.



Data Mining es un campo multi-disciplinario

11.4.1 ¿QUÉ ES EL APRENDIZAJE?

Además de definir el aprendizaje operacionalmente, necesitamos otros dos conceptos: una adecuada "tarea" a ser llevada a cabo por otro bien o mal y un aprendizaje "asunto" que lleva a cabo la tarea. Nuestra simple definición de aprendizaje es:

"Un aprendizaje individual es como llevar a cabo una tarea adecuadamente para hacer una transición de una situación en la cual la tarea no pueda ser llevada a situaciones en las cuales la misma tarea pueda ser realizada bajo las mismas circunstancias."

11.4.2 SISTEMAS QUE APRENDEN POR SI MISMOS

Ya hemos planteado que la computadora tiene la capacidad de aprender. Si una computadora, cuando es instruida, la primera vez no puede llevar a cabo una tarea particular y mas tarde, bajo las mismas circunstancias, ésta puede, decimos que ha aprendido algo. Esta conclusión es un tanto vaga, puesto que nosotros somos quien hacemos los programas. Si queremos que una computadora resuelva ecuaciones

diferenciales, simplemente escribimos un programa que sea capaz de resolverlo. Así, la computadora ha aprendido algo, aunque no podemos verdaderamente hablar de una computadora que aprenda por sí misma.

Con esto, llegamos a la conclusión y a la siguiente definición:

Una computadora que aprende por sí misma puede generar programas por si misma, y haciendo posible esto, lleva a cabo nuevas tareas.

11.4.3 CONCEPTO DE APRENDIZAJE

Hay una variedad de técnicas para hacer posible que las computadoras aprendan conceptos. Una muy importante cualidad de buenos algoritmos de aprendizaje es que aprendan definiciones consistentes y completas. Una definición de un concepto es completa si ésta reconoce todas las instancias de un concepto.

Una definición de un concepto es consistente si ésta no clasifica algún ejemplo negativo que caiga bajo el concepto.

Un muy importante elemento en el aprendizaje de una máquina es el lenguaje en el cual expresamos la hipótesis describiendo el concepto. Este lenguaje podría ser un lenguaje de computadora especializado como PROLOG o LISP, o una forma especial de representación del conocimiento usando tablas de base de datos.

Algunos aspectos que son importantes cuando describimos algoritmos de aprendizaje:

a) **Algoritmos supervisados contra no supervisados:** Algunos algoritmos necesitan el control de un humano durante su ejecución; cada algoritmo es llamado supervisor. Otros algoritmos pueden operar sin la interacción humana. Ahí encontramos algunas formas de supervisión. Puede ser que el operador humano controla sólo aquellos parámetros que influencían el rendimientos de los algoritmos durante su ejecución, pero también es posible que el operador tenga una extensa interacción con el algoritmo y necesita proveer todos los tipos de ejemplos o información adicional.

b) **Conocimiento de la experiencia:** En muchas situaciones de aprendizaje, no es realista el presuponer que un algoritmo puede aprender sin algún conocimiento experimentado, y aunque esto no es estrictamente necesario, el conocimiento experimentado será en muchos casos acelerar el proceso de aprendizaje considerablemente. Este conocimiento profundo es provisto en la mayoría de los casos dependiendo en la forma de la representación del conocimiento que usemos. Por ejemplo, en la programación lógica inductiva el conocimiento profundo podría ser un programa en PROLOG que contenga claúsulas describiendo aquellos aspectos de las teorías que son conocidas. En el caso de la inducción de árboles de decisión, el conocimiento profundo podría consistir de partes del árbol de decisión ya aprendido en ocasiones anteriores.

c) **Tendencia:** Es un mecanismo empleado por un sistema de aprendizaje que obliga la búsqueda de una hipótesis. Para cualquier algoritmo de aprendizaje hay una rica variedad de tendencias. La búsqueda de una tendencia en cuáles regiones de esta búsqueda del espacio de la solución es menos probable que encontremos una solución.

d) **Aprendizaje en conjunto contra aprendizaje incremental:** Un algoritmo de aprendizaje en conjunto lleva todos los datos al mismo tiempo y tratan de crear una hipótesis basada en sus datos. Un algoritmo de aprendizaje incremental lleva una nueva pieza de información en cada ciclo de aprendizaje y lleva a revisar la teoría usando los datos nuevos. Para el aprendizaje en conjunto esto no es necesario; por eso en la mayoría de los casos resultan más fáciles que en el aprendizaje incremental. Muchos algoritmos usados en Data Mining son del tipo del aprendizaje en conjunto.

e) **Ruido y redundancia:** Es una conexión entre redundacia y la noción del ruido. Ruido es la perturbación aleatoria de una señal transmitida; en el contexto de KDD y Data Mining esto significa errores en una tabla de la base de datos. La redundancia se refiere a los elementos de un mensaje que pueden ser generados de otras partes del mismo mensaje.

11.4.4 EL PROCESO DE DESCUBRIMIENTO DEL CONOCIMIENTO.

Muchas veces los pasos que constituyen el proceso de KDD no están tan claramente diferenciados. Las interacciones entre las decisiones tomadas en diferentes pasos, así como los parámetros de los métodos utilizados y la forma de representar el problema suelen ser extremadamente complejos. Pequeños cambios en una parte pueden afectar fuertemente al resto del proceso.

Sin quitar importancia a ninguno de estos pasos del proceso de KDD, se puede decir que el Data Mining es la parte fundamental, en la que más esfuerzos se han realizado.

Históricamente, el desarrollo de la estadística nos ha proporcionado métodos para analizar datos y encontrar correlaciones y dependencias entre ellos. Sin embargo, el análisis de datos ha cambiado recientemente y ha adquirido una mayor importancia, debido principalmente a tres factores:

1. Incremento de la potencia de los ordenadores. Aunque la mayoría de los métodos matemáticos fueron desarrollados durante los años 60 y 70, la potencia de cálculo de los grandes ordenadores de aquella época (equivalente a la de los ordenadores personales de hoy en día) restringía su aplicación a pequeños ejemplos "de juguete", fuera de los cuales los resultados resultaban

demasiado pobres. Algo similar ha ocurrido con la capacidad de almacenamiento de los datos y su coste asociado.

2. Incremento del ritmo de adquisición de datos. El crecimiento de la cantidad de datos almacenados se ve favorecido no sólo por el abaratamiento de los discos y sistemas de almacenamiento masivo, sino también por la automatización de muchos experimentos y técnicas de recogida de datos. Se estima que la cantidad de información almacenada en todo el mundo se duplica cada 20 meses; el número y tamaño de las bases de datos probablemente crece más rápidamente.

3. Por último, han surgido nuevos métodos, principalmente de aprendizaje y representación de conocimiento, desarrollados por la comunidad de inteligencia artificial, estadística y física de dinámicas no lineales. Estos métodos complementan a las tradicionales técnicas estadísticas en el sentido de que son capaces de inducir relaciones cualitativas generales, o leyes, previamente desconocidas.

Estos nuevos métodos matemáticos y técnicas software, para análisis inteligente de datos y búsqueda de regularidades en los mismos, se denominan actualmente técnicas de minería de datos o Data Mining. A su vez, el Data Mining ha permitido el rápido

desarrollo de lo que se conoce como descubrimiento de conocimiento en bases de datos.

Las técnicas de Data Mining han surgido a partir de sistemas de aprendizaje inductivo en ordenadores, siendo la principal diferencia entre ellos los datos sobre los que se realiza la búsqueda de nuevo conocimiento. En el caso tradicional de aprendizaje en ordenadores (machine learning), se usa un conjunto de datos pequeño y cuidadosamente seleccionado para entrenar al sistema. Por el contrario, en la minería de datos se parte de una base de datos, generalmente grande, en la que los datos han sido generados y almacenados para propósitos diferentes del aprendizaje con los mismos. (Gómez, 1998)

El proceso del descubrimiento del conocimiento consiste en seis escenarios:

1. La selección de datos
2. Limpieza
3. Enriquecimiento
4. Codificación
5. Data Mining
6. Reporteo

El quinto escenario, minería de datos, es la fase del descubrimiento real. A través de la metodología, como es presentada, se dá la impresión de que se tiene una trayectoria lineal a través del proceso, donde entra a la izquierda, viaja a la derecha y sale , este no es el caso. En cada escenario, la mina de datos puede regresar una o más fases; por ejemplo, cuando en la codificación o las fases de los datos, el Data Mining puede hacer que la fase de limpieza sea incompleta, o puede descubrir nuevos datos y usarlos para enriquecer otros datos existentes.

En una situación óptima, el Data Mining es un proceso que está realizándose. Las organizaciones deberían continuamente trabajar sobre sus datos, constantemente identificando nueva información necesaria y tratando de mejorar los datos para hacer esto las mejores metas. En esta forma, una organización llegará a ser un sistema de aprendizaje. Muchas de las fases necesitan la entrada de una gran distribución de creatividad, con la cual un proceso habilita y fomenta su creatividad rechazando la imposición de cualquier límite en posibles actividades.

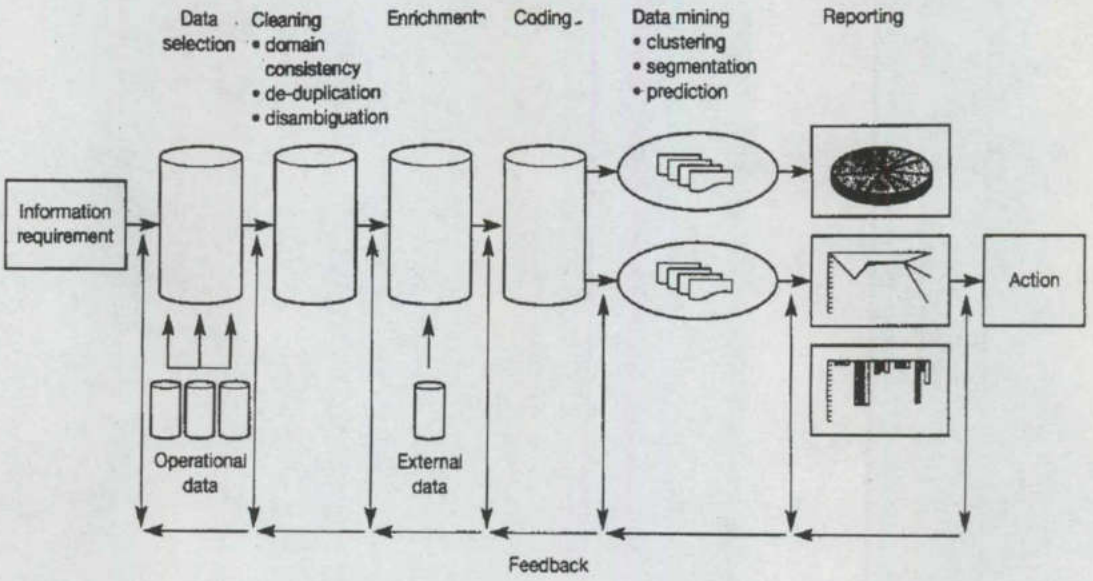


Diagrama del proceso de descubrimiento del conocimiento

11.5 EL PROCESO DEL DESCUBRIMIENTO DEL CONOCIMIENTO EN DETALLE.

Usaremos un ejemplo consistente, se trata de una base de datos de un editor de revista. El editor vende cinco tipos de revistas: carros, casas deportes, música y comics. El objeto del proceso de Data Mining es encontrar nuevos, grupos de clientes a fin de instalar un ejercicio de mercadeo. Por lo tanto, estamos interesados en preguntar: ¿Cuál es el típico perfil de un lector de una revista de autos? y ¿Existe una correlación entre un interés en carros y un interés en comics?. Discutiremos este problema usando un número de pequeñas bases de datos con cerca de 1000 registros, a través de las cuales muchas de estas son llamadas una genuina aplicación de KDD, trabaja bien por ilustración. Casi todas las técnicas son discutidas a escala, considerando que las bases de datos contienen millones de registros.

11.5.1 SELECCIÓN DE DATOS

Esta es una selección de datos operacionales del sistema de facturación de la editorial y contiene información de la gente que se ha suscrito a la revista. Los registros consisten de: número de cliente, nombre, dirección, fecha de suscripción y tipo de revista. Además de facilitar el proceso KDD, una copia de estos datos operacionales es tomada y depositada en una base de datos separada.

Número del cliente	Nombre	Dirección	Fecha de compra	Revista comprada
23003	Johnson	1 Downing Street	04-15-94	carro
23003	Johnson	1 Downing Street	06-21-93	música
23003	Johnson	1 Downing Street	05-30-92	comic
23009	Clinton	2 Boulevard	01-01-01	comic
23013	King	3 High Road	02-30-95	deportes
23019	Jonson	1 Downing Street	01-01-01	casa

Datos originales

11.5.2 LIMPIEZA

Existen algunos tipos de procesos de limpieza, de los cuales pueden ser ejecutados en progreso mientras otros son invocados sólo después de que la contaminación es detectada en la codificación o en el descubrimiento del escenario.

Un muy importante elemento en la operación de limpieza es la de-duplicación de registros. En una base de datos normal, algunos clientes serán representados por algunos registros, aunque en algunos casos esta será resultado de negligencia, que como personas tenemos errores al escribir, o de clientes mudándose de un lugar a otro sin notificar el cambio de domicilio. También algunos casos en los cuales la

gente deliberadamente deletrean sus nombres incorrectamente o dan información incorrecta acerca de su persona, especialmente en situaciones donde los individuos se les ha negado algún tipo de seguro. Por supuesto que es importante para cualquier compañía ser consciente de las anormalidades de la base de datos. Aunque el Data Mining y la limpieza de los datos son dos diferentes disciplinas, tienen mucho en común y los algoritmos de reconocimiento de patrones pueden ser aplicados en la limpieza de datos.

En el ejemplo tenemos al Sr. Johnson y al Sr. Jonson en la base de datos. Tienen diferente número de cliente pero la misma dirección, lo cual es una fuerte indicación de que son la misma persona pero que una de ellas es incorrecta. Por supuesto, nunca podemos estar seguros de esto, pero un algoritmo de de-duplicación usando técnicas de análisis de patrones podrían identificar la situación y presentar esto al usuario para tomar una decisión. Este tipo de contaminación ocurre frecuentemente en las bases de datos.

El segundo tipo de contaminación que frecuentemente ocurre es la falta de consistencia de dominio. Note que la tabla originados registros fechados el 1ro. de enero de 1901, aunque la compañía probablemente no existía en aquel entonces. Este tipo de contaminación es dañina, porque es difícil de rastrear, pero será grandiosa influencia el tipo de patrones que encuentre cuando aplique Data Mining en esta tabla. En algunas bases de datos, el análisis muestra un inesperado alto número de

personas nacidas el 11 de noviembre. Cuando la gente está forzada a llenar su fecha de nacimiento en una pantalla y ellos no saben o no quiere divulgarlo, se inclinan a escribir 11-11-11. Esto es desastroso en el contexto del Data Mining, si la información es desconocida debería ser representada como tal en la base de datos. En el ejemplo, reemplazamos parte de los datos con valores nulos y corregidos otros dominios inconsistentes.

Número del cliente	Nombre	Dirección	Fecha de compra	Revista comprada
23003	Johnson	1 Downing Street	04-15-94	Carro
23003	Johnson	1 Downing Street	06-21-93	música
23003	Johnson	1 Downing Street	05-30-92	comic
23009	Clinton	2 Boulevard	NULO	comic
23013	King	3 High Road	02-30-95	deportes
23003	Johnson	1 Downing Street	12-20-94	casa

11.5.3 ENRIQUECIMIENTO

En el ejemplo, se supone que se ha obtenido información extra acerca de nuestros clientes que consisten en la fecha de nacimiento, ingresos, tarjeta de crédito, e información acerca de sus propiedades como si tiene carro o casa propia.

Para este ejemplo, no es particularmente importante saber cómo la información fué recopilada, pero es necesario apreciar que la información nueva puede agregarse fácilmente a los registros de los clientes existentes.

Nombre del Cliente	Fecha de Nacimiento	Ingresos	Crédito	Carro Propio	Casa propia
Johnson	04-13-76	$18,500	$17,800	No	no
Clinton	10-20-71	$36,000	$26,600	Si	no

11.5.4 CODIFICACIÓN

Los datos del ejemplo pueden someterse a un número de transformaciones. Primero, la información extra que fué obtenida para enriquecer nuestra base de datos es agregada a los registros describiendo a los individuos.

En el siguiente escenario, seleccionamos sólo aquellos registros que tengan suficiente información valiosa. Aunque es difícil de dar reglas detalladas de este tipo de operación, esta es una situación que puede ocurrir frecuentemente en la práctica. En muchas tablas que son recolectadas de datos operacionales, muchos de los datos importantes se pierden y es imposible recuperarlos. Por lo que tiene que tomar una decisión deliberada, pero no pasarlo por alto ó definitivamente borrarlo.

Una regla general dice que cualquier borrado de datos debe ser una decisión consciente, después un completo análisis de las posibles consecuencias. En algunos casos, en especial en detección de fraudes, la falta de información puede ser un indicador valioso de patrones interesantes.

La fase de codificación ha consistido de ninguna otras cosas mas que de simples operaciones SQL pero ahora estamos entrando a una escenario donde seremos capaces de rendir mas transformaciones creativas sobre los datos. En este momento, la información en nuestras bases de datos es mucho muy detallada para ser usadas como entrada para algoritmos de reconocimiento de patrones. Tome por ejemplo, la

noción de una fecha de nacimiento: un algoritmo que introduce gente con la misma fecha de nacimiento en una clase de cliente seguro es obviamente mucho más detallada para reconocimiento de patrones y en este caso, necesitamos direcciones recodificadas dentro de códigos regionales. La forma en que será nuestro código de información, en una gran extensión, determina el tipo de patrones que encontraremos. Codificación, por lo tanto, es una actividad creativa que tiene que ser desempeñada repetidamente a fin de llevar a los mejores resultados.

En lugar de hacer un análisis de tiempo, nuestro editor está más interesado en las relaciones entre lectores de las diferentes revistas. Esto significa que no será investigadas las conexiones entre un producto y fechas de suscripción, pero si entre clases de productos, así que las fechas de suscripción son las menos importantes por el momento.

Esta es la nueva tabla que resultó del proceso de codificación. Una tabla en este formato, como todo, no es de mucha ayuda si se quiere encontrar relaciones entre las diferentes revistas. Cada suscripción es representada por un registro, aunque éste debería de ser más eficiente al tener una vista global de todas las revistas a la que cada lector se suscribió.

Num. Cliente	Edad	Ingresos	Crédito	Carro propio	Casa propia	Región	Mes de compra	Revista comprada
23003	20	18.5	17.8	0	0	1	52	carro
23003	20	18.5	17.8	0	0	1	42	música
23003	20	18.5	17.8	0	0	1	29	comic
23009	25	36.0	26.6	1	0	1	Nulo	comic
23003	20	18.5	17.8	0	0	1	48	casa

Proporcionamos una transformación final en la tabla y creamos sólo un registro para cada lector. En lugar de tener un atributo "revistas" con cinco diferentes posibles valores, creamos cinco atributos binarios, uno para cada revista.

Si el valor del atributo es 1, esto significa que el lector es un suscriptor, de otra forma, el valor es 0. Cada operación es llamada "aplanando" - un atributo con cardinalidad n es reemplazado por n atributos binarios. Esta es una operación de codificación que ocurre frecuentemente en un contexto KDD.

Ahora tendremos finalmente codificado nuestros datos colocados en la forma: número de cliente, edad, ingresos, crédito, información concerniente a auto y casa propia, código de área, y cinco bits indicando en qué revistas el cliente se ha suscrito. Esta es una buena base con la cual comenzamos el proceso real de Data Mining.

Revistas compradas:

Núm. Cliente	Edad	Ing.	Crédito	Carro propio	Casa propia	Región	Revista de carros	Rev. de casas	Rev. Dep.	Rev. de música	Rev. de comic
23003	20	18.5	17.8	0	0	1	1	1	0	1	1
23009	25	36.0	26.6	1	0	1	0	0	0	0	1

11.5.5 DATA MINING

El descubrimiento de escenarios del proceso KDD es fascinante. El Data Mining no es una simple técnica con la idea de que es mucho conocimiento escondido en los datos que se muestran en la superficie.

Cualquier técnica que ayuda a extraer mas de sus datos es útil, así que, las técnicas de Data Mining forman del todo un grupo heterogéneo. Aunque diferentes **técnicas** son usadas para propósitos diferentes, aquellas que son de interés en el presente contexto son:

- **Herramientas de consulta**
- **Técnicas de análisis estadístico**
- **Visualización**
- **Procesamiento analítico en línea (OLAP)**
- **Aprendizaje basado en casos**
- **Arboles de decisión**
- **Reglas de asociación**
- **Redes neuronales**
- **Algoritmos genéticos (Adrians, 1996)**

La tecnología fundamental en el descubrimiento de conocimientos son los algoritmos para patrones y relaciones. Muchos de éstos algoritmos se derivaron de las actividades de investigación en inteligencia artifcial y aprendizaje con máquinas. **La tecnología se describe desde la perspectiva de categorías genéricas**:

Análisis de Dependencias.

Los algoritmos de análisis de dependencias extraen las dependencias entre elementos u objetos en el Data Warehouse. Una dependencia es interesante ya que revela dependencias desconocidas entre los datos, y es posible que se describa la relación casual entre los elementos de datos de interés. Por lo tanto, estos algoritmos se utilizan para predecir el valor de un objeto de datos a partir del valor de otro. Por lo regular, las dependencias no definen una relación exacta o segura, sino más bien, un valor de probabilidad, el valor de confianza.

Clasificación.

Estos algoritmos agrupan los datos en clases significativas. Por ejemplo, se emplea el algoritmo para crear un perfil de clientes preferenciales, o se usa para generar una base a fin de detectar desviaciones. Los algoritmos de agrupamiento se usan para descubrir clases de forma automática. Algunos de los algoritmos comunes de

agrupamiento son de reconocimiento de patrones, la generación de perfiles, el agrupamiento lineal y el agrupamiento conceptual. La clasificación mejora si el sistema de descubrimiento de conocimientos utiliza al analista empresarial en el ciclo. En este escenario, el poder computacional del sistema de descubrimiento de conocimientos se combinan con el conocimiento de la materia y la habilidad visual del analista empresarial.

Descripción de conceptos: Estas técnicas resumen las reglas de interés o la descripción de la clase. Esta categoría emplea los dos tipos de descripciones siguientes.

- Característica: describe qué tienen en común los elementos de datos de la clase
- Discriminatoria: describen en qué difieren dos o más clases.

Los algoritmos comunes para describir conceptos incluyen los árboles de decisión, los inductores de árboles de decisión, las redes neuronales y los algoritmos genéticos.

11.5.6 REPORTEO

El reporteo se realiza a través de las técnicas de visualización descritas en hojas posteriores.

12. DESCRIPCIÓN DE LAS TÉCNICAS MÁS COMÚNMENTE USADAS EN DATA MINING

12.1 ANÁLISIS PRELIMINAR DE LOS DATOS USANDO HERRAMIENTAS DE CONSULTA

El primer paso en un proyecto de Data Mining debería ser siempre un análisis áspero de los datos usando herramientas de consulta tradicionales. Aplicando sólo SQL, puede obtener abundancia de información. Sin embargo, antes de aplicar mas algoritmos de análisis de patrones avanzados, se necesita conocer algunos aspectos básicos y estructuras de datos. Con SQL podemos descubrir sólo datos superficiales, la cual es información que es fácilmente accesible de los datos; aunque no podemos encontrar datos ocultos, el 80% de la información interesante puede ser abstraída de una base de datos usando SQL. El 20% restante, de la información oculta requiere más técnicas avanzadas y para grandes organizaciones, este 20% puede comprobar vital importancia. (Adrians, 1996)

12.2 ANÁLISIS ESTADÍSTICO

Los sistemas de análisis estadístico, también conocido como análisis de datos, se usan para detectar patrones no usuales de datos. Algunas de las técnicas de modelado estadístico y matemático que se emplean son el análisis lineal y no lineal, el análisis de regresión continua y logística, el análisis de univariación y multivariación y el análisis de series históricas.

Las herramientas de análisis estadístico se utilizan en diversas aplicaciones empresariales: incrementar la participación en el mercado y las utilidades, detectando las mejores oportunidades, aumentar la satisfacción del usuario por medio del mejoramiento de la calidad en productos y servicios, e impulsar los márgenes a través de la modernización de la manufactura de los productos y de la logística. Las herramientas de análisis estadístico existen desde hace tiempo y son las herramientas más desarrolladas que se tienen para el Data Mining. Han servido para reducir el tiempo de análisis, lo cual libera los recursos limitados para otras actividades de análisis, lo que a su vez conduce a una mejor toma de decisiones.

12.2.1 USO DE LAS HERRAMIENTAS DE ANÁLISIS ESTADÍSTICO

Para utilizar una herramienta de análisis estadístico, los usuarios empresariales deben seleccionar y extraer los datos adecuados del Data Warehouse.

A continuación, los usuarios empresariales deben invocar las funciones de visualización y analítica, disponibles en la herramienta de análisis estadístico, para descubrir relaciones entre los datos y construir modelos estadísticos y matemáticos a fin de interpretar los datos. Se usa un proceso interactivo e iterativo para refinar el modelo; la meta consiste en desarrollar el modelo de mejor ajuste para convertir los datos en información. Los analistas empresariales con capacidad para resolver problemas y experiencia en su campo son fundamentales para seleccionar el modelo que se ajusta mejor.

12.2.2 CARACTERÍSTICAS DE LAS HERRAMIENTAS DE ANÁLISIS ESTADÍSTICO.

Dada la complejidad de muchas de las tareas del análisis estadístico, las herramientas para el efecto deben ofrecer lo siguiente:

12.3 FUNCIONES DE VISUALIZACIÓN

Estas funciones ayudan a descubrir relaciones entre grandes cantidades de datos. Por ejemplo, las funciones deben reconocer patrones en las series históricas de datos y exhibir gráficas de línea o logarítmicas, o realizar ajustes de curva para encontrar en los datos la "regla o patrón empresarial", o manipular los datos mediante el agrupamiento automático de valores de variable únicas seleccionadas, o alterando el punto de inicio y el tamaño de los histogramas.

12.3.1 FUNCIONES EXPLORATORIAS

Estas funciones ayudan a elegir la función estadística y el modelo correctos que se ajustan a los datos. Algunas de estas funciones son tablas multidimensionales de pivoteo, ayuda orientada al análisis, e identificación de valores extremos y distantes. La herramienta debe producir y presentar cuadros, gráficas y tablas para el analista empresarial, en forma dinámica y automática, como parte del proceso de exploración.

12.3.2 FUNCIONES Y EXPLORACIONES ESTADÍSTICAS

Estas funciones y operaciones ofrecen un rico conjunto de herramientas, tales como el análisis de regresión, tanto continúo como logístico; el análisis de las series históricas, incluyendo autocorrelación; transformaciones rápidas de Fourier y pronósticos; análisis de variación múltiple; ANOVA; CHAID; pruebas no paramétricas y análisis de respuesta múltiple.

12.3.3 FUNCIONES DE ADMINISTRACIÓN DE DATOS

Estas funciones profundizan al detalle, examinan subconjuntos de datos, discriminan valores extremos, comparan subconjuntos, etc.

12.3.4 FUNCIONES DE GRABACIÓN Y REPRODUCCIÓN

Estas funciones graban los pasos del análisis, transfieren los registros a otro analista empresarial, y reproducen luego la tarea completa de análisis. Las funciones de grabación deben incluir los pasos del análisis, el proceso de selección de conjuntos de datos, una paleta o carrusel de cuadros y gráficas seleccionados y cualquier otra información que se vaya a comunicar. Esta es la clave para comunicar y compartir, tanto los resultados de la tarea de análisis estadístico, como las técnicas analíticas y el proceso aplicados.

12.3.5 HERRAMIENTAS DE PRESENTACIÓN

Estas herramientas comunican los datos complejos y el análisis de cuadros, gráficas y tablas sencillas. La herramienta debe convertir con rapidez los datos de un tipo de cuadro y, cuando se necesite, exhibirlos en un tipo de cuadro diferente. La herramienta debe también mostrar la variedad de tipos de cuadros, gráficas y tablas a los que se ajustan los datos, de modo que se seleccionen con facilidad las mejores opciones de presentación.

Un conjunto básico de los cuadros y gráficos requeridos consiste en gráficas lineales x-y y de dispersión, gráficas de cuadro, histogramas, gráficas de barra (de pastel y de área), de intervalos, de superficie tridimensional y contorno, cuadros estadísticos (como de Pareto y de barra X) y reportes tales como tabulaciones cruzadas.

12.3.6 JUEGO DE HERRAMIENTAS DEL DESARROLLADOR

Utilice este juego para enlazarse con facilidad a aplicaciones de escritorio y complementar componentes para el análisis estadístico y la elaboración de cuadros, gráficas y reportes. La disponibilidad de un lenguaje de programación orientado a objetos con una interfaz tipo apuntar y hacer clic, así como el intercambio de datos mediante técnicas como OLE (circulación e incorporación de objetos), reforzará al analista empresarial para incluir el análisis estadístico en aplicaciones de escritorio para soporte de decisiones.

12.3.7 TIEMPO DE RESPUESTA RESPONSABLE

Este periódo que se mide en minutos o incluso en horas, es aceptable para algunas decisiones empresariales. Por siempre, existen excepciones, como en la industria de seguros; el tiempo de respuesta en días es inaceptable ya que la relevancia del análisis declina al desactualizarse los datos, y la oportunidad se aleja.

12.3.8 RETOS EN LA APLICACIÓN DE LAS TÉCNICAS DE ANÁLISIS ESTADÍSTICO

El análisis estadístico es una técnica poderosa para comprender a los clientes, mercados, productos y otros parámetros empresariales importantes. Pero existen algunos retos, como los siguientes:

- Representa un trabajo intenso
- El éxito depende en gran medida de la habilidad del analista empresarial para solucionar problemas.
- Muchas veces, el analista empresarial no sabe qué buscar, o no puede seleccionar variables separadas para iniciar el proceso de análisis.
- Resulta difícil integrar y analizar datos no numéricos en un esfuerzo por segmentar el mercado.
- Es arduo alcanzar un tiempo de respuesta aceptable a un costo razonable.

12.4 TÉCNICAS DE VISUALIZACIÓN

Son un método muy útil del descubrimiento de patrones en un conjunto de datos, y pueden ser usados al principio del proceso de Data Mining al llevar al descubrimiento de la calidad de los datos y donde los patrones son encontrados. Las posibilidades interesantes son ofrecidas por herramientas tridimensionales orientadas a objetos tal como un inventor, el cual habilita al usuario a explorar estructuras interactivamente tridimensionales. Cada técnica es desarrollada rápidamente: técnicas gráficas avanzadas en realidad virtual facilita a la gente a errar a través de espacios de datos artificiales, mientras el desarrollo histórico de datos puede ser desplegado como un tipo de película animada. Para muchos usuarios, sin embargo, rasgos avanzados no son accesibles y tienen que contar con simples técnicas de desplegado de gráficas que están contenidas en las herramientas de consulta ó herramientas de Data Mining. Estos métodos simples pueden proveernos de una riqueza de información. Una técnica elemental que puede ser de gran valor es el llamado diagrama de dispersión; en ésta técnica, la información sobre dos atributos es desplegada en un espacio cartesiano. Los diagramas de dispersión pueden ser usados para identificar interesantes subconjuntos de datos colocados de esta forma, podemos enfocar el resto del proceso del Data Mining. Es un campo integro de investigación dedicado a la búsqueda de interesantes proyecciones de datos, llamado también proyección de búsqueda.

En el ejemplo, se ha hecho una proyección a lo largo de dos dimensiones: ingresos y edades. Vemos que en promedio, la gente joven con un bajo ingreso tiende a leer la revista de música.

Ahora comparamos cómo tan simples técnicas de visualización pueden ayudar a dar una noción de la estructura de los datos. Las mejores formas de explorar un dato es a través de un ambiente tridimensional interactivo.

12.4.1 PROBABILIDAD Y DISTANCIA

Estas son otras razones para concebir registros como puntos en un espacio de datos multidimensional. La metáfora del espacio es muy útil en un contexto de Data Mining. Usando ésta metáfora podemos determinar a distancia entre dos registros en su espacio de datos: registros que están cerca y cada uno son muy distintos, y registros que son muy lejos de ser removidos de otras representaciones individuales que tienen poco en común. En el ejemplo, la base de datos contiene atributos como la edad, ingresos y crédito; éstos tres atributos forman un espacio de datos tridimensional y podemos analizar las distancias entre registros en este espacio. Otra ventaja de una buena codificación viene a dislumbrar - además de lograr una buena comparación entre valores, debemos normalizar los atributos. La edad, por ejemplo,

en rangos de 1 hasta 100 años, mientras los ingresos tienen un rango de 0 a aproximadamente 100,000 dólares al mes. Ahora, si usamos estos datos sin corrección, los ingresos serán por supuesto, un atributo mucho más distintivo que la edad, y no es lo que queremos. Por eso, dividimos los ingresos por 1000, además de obtener una medida que tiene el mismo orden de magnitud como la edad. Hacemos lo mismo para el atributo crédito. Si escalamos todos los atributos al mismo orden de magnitud obtenemos una confiable distancia medida entre los diferentes registros. En el ejemplo, usando la medida de la distancia euclidiana, la distancia entre el cliente 1 y el cliente 2 es 15.

En esta forma, los registros se convierten en puntos en un espacio de datos multidimensional. Para los espacios de los datos con baja dimensionalidad es fácil de visualizar nubes de datos, y de algunos podemos identificar interesantes grupos meramente por inspección visual. En muchos casos, sin embargo, necesitamos mas programas de búsqueda avanzados para descubrir cada grupo, y predicciones interesantes pueden también ser visualizadas en esta forma: en algunas es posible identificar un grupo visual de clientes potenciales que están muy contentos de comprar un producto acertado. En el ejemplo de datos se contempló la edad, los ingresos y la forma de crédito en un espacio tridimensional ideal en el cual se realiza este tipo de análisis de agrupación.

12.5 HERRAMIENTAS OLAP

El término OLAP (On Line Analytical Processing) fue acuñado por Codd y asociados en un artículo titulado "Providing OLAP to user-analysts: An IT mandate" publicado en 1993 y que fue apoyado por Arbol Software Corporation, los creadores y vendedores de ESSBASE una de las primeras herramientas OLAP que aparecen en el mercado. (Gómez, 1998)

12.5.1 ANÁLISIS MULTIDIMENSIONAL.

En el análisis multidimensional, los datos se representan mediante dimensiones como producto, territorio y cliente. Por lo regular las dimensiones se relacionan en jerarquías, por ejemplo, ciudad, estado, región, país y continente, o estado, territorio y región. El tiempo es también una dimensión estándar con su propia jerarquía como día, semana, mes, trimestre, año, ó día y año calendario. Para facilitar el análisis complejo, el procesamiento analítico o análisis multidimensional presenta una visión empresarial sencilla de los datos. Un usuario empresarial puede acceder los ingresos por departamento y tienda para los últimos cuatro trimestres, para un conjunto dado de productos. Los resultados se pueden pivotear o girar para cambiar los ejes y la

perspectiva, además, los usuarios empresariales pueden navegar por las dimensiones profundizando u obteniendo resúmenes a lo largo de los elementos de una dimensión, o penetran a través de las dimensiones para ver otras perspectivas. Un equipo computarizado ofrece también capacidades como posiciones al principio o al final de la lista; promedios móviles, tasas de crecimiento, cálculos financieros de interés, tasas internas de recuperación y de depreciación y conversiones monetarias y funciones estadísticas.

Al procesamiento analítico o análisis multidimensional se le conoce también como procesamiento analítico en línea (OLAP). Se apoya en una visión multidimensional de los datos empresariales en el Data Warehouse y, puede tener un motor de depósito de base de datos multidimensional.

12.5.2 PROCESAMIENTO ANALÍTICO EN LÍNEA (OLAP)

En un Data Warehouse, se depositan datos para consulta, análisis y divulgación, a diferencia del procesamiento de transacciones en línea (OLTP) por las siglas de Online Transaction Processing, en donde los datos se reúnen y almacenan para operación y control. OLTP es una tecnología de procesamiento analítica que crea

nueva información empresarial a partir de los datos existentes, por medio de un rico conjunto de transformaciones empresariales y cálculos numéricos.

12.5.2.1 DEFINICIÓN DE OLAP.

El procesamiento analítico en línea, es una tecnología de análisis de datos que hace lo siguiente:

- Presenta una visión multidimensional lógica de los datos en la Data Warehouse. La visión es independiente de como se almacenan los datos.
- Comprende siempre la consulta interactiva y el análisis de datos. Por lo regular, la interacción de varias pasadas, lo cual incluye la profundización en niveles cada vez más detallados o el ascenso a niveles superiores de resúmen y adición.
- Ofrece opciones de modelado analítico, incluyendo un motor de cálculo para obtener proporciones, desviaciones, etc., que comprende mediciones de datos numéricos a través de muchas dimensiones.
- Crea resúmenes y adiciones (también conocidas como consolidaciones), jerarquías, y cuestiona todos los niveles de adición y resúmen en cada intersección de las dimensiones.

- Maneja modelos funcionales de pronóstico, análisis de tendencias y análisis estadísticos.
- Responde con rapidez a las consultas, de modo que el proceso de análisis no se interrumpe y la información no se desactualiza.
- Tiene un motor de depósito de datos multidimensional que almacena los datos en arreglos. Estos arreglos son una presentación lógica de las dimensiones empresariales. (Harjinder, 1996)

Los sistemas OLAP pertenecen a los sistemas de información para ejecutivos, EIS, utilizados para proporcionar al nivel estratégico información fiable sobre los indicadores clave del funcionamiento de una organización.

Mecanismos que proporcionan un rápido análisis de información multidimensional compartida; hay muchas áreas en las que esto es sumamente importante, como por ejemplo marketing y análisis de ventas. Y es para estas áreas para las que se desarrollaron los primeros sistemas OLAP.

Rubenstrunk **divide las herramientas OLAP en tres categorías** basándose en la profundidad de sus capacidades analíticas:

1. En el primer nivel se encuentran aquellas herramientas que proporcionan una visión multidimensional, informes y un análisis pequeño y simple.
2. El nivel siguiente añade un mejor análisis.
3. Y en el último nivel nos podemos encontrar con algoritmos de búsqueda, análisis de patrones, reglas de localización, etc.

Otra característica que varía de una aplicación a otra es la necesidad de interacción; si los usuarios realizan preguntas bien formuladas, no tendrán que esperar mucho a las respuestas. También podrán dedicarse a explorar los datos cuando no saben exactamente lo que se busca, como otro método de interacción que les permita extraer alguna conclusión.

Finalmente hay que tener en cuenta que OLAP no es una herramienta de usuario sino un entorno de desarrollo de aplicaciones destinadas a solventar un abanico de necesidades de información y poner los medios para obtener una información depurada y ajustada a nuestras necesidades.

12.5.3 VENTAJAS Y DESVENTAJAS DE LA TECNOLOGÍA OLAP

La tecnología OLAP permite una mayor rapidez y efectividad a la hora del análisis de unos datos, que pueden ser de cualquier tipo, desde datos de ventas hasta datos del

alumnado de una universidad. Esta rapidez y efectividad se traduce en una información mejor y más depurada ya que mientras que sin utilizar una herramienta OLAP nos tendremos que limitar a observar exhaustivamente los datos y obtener cuatro o cinco gráficas de aquellos datos que parezcan más interesantes, con una herramienta OLAP podremos, en el mismo tiempo, comparar multitud de datos y gráficas y extraer relaciones que a simple vista nunca se habrían ocurrido.

Se ha buscado algún otro tipo de tecnología que se encuentre compitiendo con la tecnología OLAP pero no se encontró nada. De hecho otras tecnologías como las bases de datos, hojas de cálculo, sistema de análisis de datos, etc están evolucionando hacia la tecnología OLAP, adaptando sus productos a las exigencias que puede tener una herramienta OLAP igual que adaptan sus productos para aplicarlos con tecnología de redes, un ejemplo de ello, es ORACLE.

Resulta difícil encontrar desventajas a esta tecnología a no ser que se consideren las herramientas OLAP como software que una empresa debe adquirir y entonces se presenten los inconvenientes relacionados con el software como puede ser el desembolso inicial para adquirir la herramienta, la necesidad de la existencia de una base de datos previa , ajena normalmente al software OLAP) y la necesaria actualización de esta herramienta. El segundo punto es quizás el aspecto menos conseguido de la tecnología OLAP: las herramientas típicas OLAP (como el POWERPLAY) necesitan la existencia de una base de datos ajena al programa OLAP para poder interactuar con ella. Otras empresas (como ORACLE) proporcionan ya en

sus bases de datos soporte para herramientas OLAP en un intento de juntar las dos tecnologías.

En cualquier caso y pese a las mínimas desventajas que puede traer consigo la adquisición de una herramienta OLAP, las ventajas que conlleva el uso de una herramienta OLAP para el análisis de datos son mucho mayores tanto en número como en importancia. (Universidad de Alicante, 1999)

12.5.4 ARQUITECTURA OLAP

Existe mucha confusión respecto a las arquitecturas OLAP, en lo referente a términos como ROLAP, HOLAP, MOLAP, DOLAP que cada día proliferan más. De hecho, existen una gran variedad de maneras de almacenar y procesar los datos OLAP. La gran mayoría de fabricantes de este tipo de productos sólo ofertan una gama limitada de estas arquitecturas y algunos de estos tienden a especializarse en una sola. De cualquier modo, son bastante pocos los productos que pueden almacenar y procesar los datos de más de un modo distinto (se pueden catalogar como arquitecturas distintas) y los fabricantes de estos tipos de productos, en general, suelen ser poco llamativos.

Aunque existe una gran diversidad a la hora de nombrar o llamar a las arquitecturas OLAP de acuerdo al modo de procesar y almacenar los datos, realmente sólo se consideran tres. En teoría, de nueve posibles arquitecturas básicas sólo seis son implementadas. (Universidad de Alicante, 1999)

La arquitectura OLAP una opción de análisis y reporte. Es un componente importante del bloque de acceso y uso de la arquitectura de referencia del Data Warehouse. Los componentes de la tecnología OLAP se capturan en sub-bloques de los bloques de acceso y uso.

El componente OLAP del sub-bloque de análisis y reporte representa las opciones de análisis y reporte de los servicios OLAP requeridos, mientras que la transformación a la estructura multidimensional, así como el acceso de los componentes del Data Warehouse, son parte del sub-bloque de acceso y recuperación. La arquitectura de referencia ofrece las siguientes opciones:

- Acceder los datos directamente desde el Data Warehouse, después transformar en una estructura multidimensional y almacenarlos en el depósito local de la estación de trabajo.
- Acceder los datos desde el Data Warehouse, para después transformarlos en una estructura multidimensional y almacenar esta en el Data Warehouse de una

depósito de datos multidimensional disponible con facilidad para el acceso y análisis multidimensional en la estación de trabajo.

- Acceso a los datos directamente desde el Data Warehouse para luego transformarlos en una visión multidimensional y presentarlos como una estructura multidimensional al usuario empresarial para su análisis y reporte en la estación de trabajo.

12.5.5 TECNICAS ASOCIADAS CON LA TECNOLOGÍA OLAP

12.5.5.1 BASES DE DATOS.

El problema de la dispersión de datos.

La explosión en las bases de datos multidimensionales es un fenómeno común pero poco comprendido. Este efecto puede provocar que muchas aplicaciones no funcionen correctamente, por lo tanto cualquiera que trabaje con una aplicación o con una base de datos multidimensional debe prever la posibilidad de este fenómeno y plantear que hacer.

Es importante, en primer lugar, considerar cuales son las causas que no provocan la explosión de las bases de datos (al contrario de lo que se podría pensar):

- Los datos dispersos: Es cierto que una mala redistribución o eliminación de los datos dispersos suele provocar una mayor ocupación del disco duro, pero la diferencia entre un buen o mal almacenamiento de los datos dispersos es sólo un factor más que no contribuye en gran medida.
- Almacenamiento de la base de datos multidimensional: Contrariamente a la propaganda que realizan muchos vendedores, la tecnología usada para almacenar los datos multidimensionales no provocan el fenómeno de la explosión de datos.
- Compresión de los datos: Algunos productos OLAP presentan la posibilidad de comprimir los datos, pero esto no soluciona el tema de la explosión de los mismos.
- Errores en el software: La explosión de los datos no tiene nada que ver con errores en el software o base de datos corruptas.

La manera más sencilla de almacenar datos dispersos es almacenar sólo celdas que contengan datos en alguna forma indexada. Pero esto conlleva sus propios problemas: el índice y las claves pueden llegar a ocupar incluso más espacio que los datos.

La mayoría de los grandes productos dividen los datos en grupos más pequeños y densos (objetos multidimensionales). Algunos lo hacen implícitamente presentando los datos al usuario en un formato de hypercubo, donde todos los datos en la aplicación aparecen como una sencilla estructura multidimensional. Otros hacen explícitamente una aproximación multicubo donde la base de datos multidimensional consiste en un número de objetos separados normalmente con diferentes dimensiones.

12.5.5.2 INTEGRACIÓN BASES DE DATOS-HERRAMIENTAS OLAP.

Las bases de datos multidimensionales suelen adquirir datos de otras fuentes como bases de datos relacionales, herramientas de escritorio u hojas de cálculo; también algunos datos son introducidos por usuarios finales. En un ROLAP los datos se encuentran fisicamente almacenados en un RDBMS (Base de datos Relacional), mientras que en un MDB (Base de Datos Multidimensional) los datos están almacenados en diferentes estructuras optimizados por un procesamiento multidimensional.

Si los datos son almacenados en un MDB normalmente ocuparán menos espacio que si se encuentran en el sistema originario del que se recogen los datos. Esto es consecuencia principalmente de las claves, las indexaciones y las estructuras dimensionales, que o no se necesitan o se optimizan para ocupar mucho menos espacio. También la dispersión de los datos se elimina de una manera más eficaz y los datos pueden ser comprimidos.

La mayoría de aplicaciones OLAP se encuentran orientadas para un uso interactivo de manera que el usuario espera conseguir una rápida respuesta a sus interrogantes, en pocos segundos. El principal coste en términos de tiempo consumido, a la hora de hacer los cálculos, no son los cálculos aritméticos sino la recuperación de los datos que afectan a los elementos que se van a recuperar.

Para conseguir una respuesta rápida todas las grandes aplicaciones multidimensionales necesitan precalcular algunos de los datos que van a utilizar en el análisis. En las bases de datos multidimensionales el almacenamiento de datos precalculados suele ser automático y transparente, mientras que en ROLAP normalmente se usan tablas resumen.

El hecho de precalcular toda la información necesaria puede llegar a ser un problema en lugar de una ventaja. El problema se encuentra en las relaciones multidimensionales que existen en todas las aplicaciones OLAP y en el hecho de que la mayoría de los datos introducidos son considerados dispersos. Esto provoca que los resultados precalculados basados en datos multidimensionales dispersos sean bastante más voluminosos de lo que sería deseable.

Con todo lo visto hasta ahora podemos decir que una consulta se basa en tres tipos de datos: datos de entrada introducidos por el usuarios, datos precalculados y cálculos en el momento.

Para evitar una explosión de datos se pueden adoptar dos principios. En primer lugar evitar precalcular cualquier objeto multidimensional que tenga más de cinco dimensiones ya que esto provocará que los datos dispersos se multipliquen. Por otra parte también puede resultar interesante reducir la dispersión de objetos de datos individuales mediante un buen diseño y utilizando una aproximación de multicubos en la que cada objeto tenga el mínimo número de dimensiones necesario.

La cantidad exacta de lo que debemos precalcular depende de varios factores entre ellos el hardware, la red, las características del software, el número de usuarios, la complejidad de los cálculos, etc. Normalmente los datos que precalcularemos serán:

- Datos lentos de calcular en tiempo de ejecución
- Datos que se pidan con cierta asiduidad
- Datos que constituyan la base para el cálculo de otros datos

12.5.5.3 HYPERCUBOS VS MULTICUBOS

Los diseñadores de productos multidimensionales utilizan varias estrategias para evitar la dispersión de los datos y agruparlos. Se puede escoger una de las dos principales maneras de presentar datos al usuario: hypercubos o multicubos; estas opciones no son visuales, sino que condicionan cómo los datos van a ser procesados y cuántos cálculos se van a realizar.

HYPERCUBOS

Algunos productos ofrecen un simple y único cubo como estructura de almacenaje de datos ofreciendo modelos más sofisticados de compresión de datos dispersos. Permiten introducir valores de datos mediante combinación de dimensiones y todas las partes del espacio de datos (cubo) tienen la misma dimensión. A esta estructura la llamaremos hypercubo sin limitar el número de dimensiones a algún valor fijo ni tener en cuenta que las dimensiones sean de igual tamaño. Este término se usaría de una forma específica, no de una forma general para identificar estructuras con más de tres dimensiones. De cualquier forma este término no hace referencia a un formato de almacenar los datos y puede ser aplicado tanto en bases de datos multidimensionales como en relacionales.

Los defensores de la estructura de hypercubo resaltan su simplicidad cara al usuario final. Arbor con Essbase y Cognos con Powerplay son dos de las compañías que utilizan esta estructura en sus aplicaciones.

Por último comentar que hay una variante de esta estructura de hypercubo: fringed hypercube (hypercubo bordeado). Se trata de un hypercubo más denso, con un pequeño número de dimensiones y a las que se puede añadir dimensiones de análisis.

MULTICUBOS

En los productos que utilizan esta estructura, la aplicación divide la base de datos en un conjunto de estructuras multidimensionales cada una de las cuales está formada por un subconjunto de las dimensiones de la base de datos. Cada producto llama de una manera a estas pequeñas estructuras: variables, estructuras, cubos, universos, etc... pero todos hacen referencia al mismo tipo de subestructura.

Los defensores de estos sistemas destacan la gran versatilidad de estos sistemas, su potencia y su eficiencia, especialmente con los datos dispersos, considerando los hypercubos como un simple subconjunto de su aproximación.

Los productos ROLAP también pueden utilizar una estructura en multicubos si son capaces de manejar múltiples tablas base, cada una con una dimensión diferente.

Se pueden identificar dos tipos principales de multicubos: el block multicubo y el multicubo series. El multicubo bloque utiliza dimensiones ortogonales así que no hay dimensiones especiales en el nivel de datos. Un bloque puede consistir en cualquier número de dimensiones definidas y tanto medidas como tiempo son tratados como dimensiones. Los multicubos serie tratan cada medida o variable como series de tiempo, con su propio conjunto de dimensiones.

De manera general los multicubos son más versátiles pero los hypercubos son más sencillos de comprender cara al usuario. Los usuarios finales suelen preferir los

hypercubos debido a su visión de alto nivel mientras que los profesionales con experiencia prefieren multicubos por su gran flexibilidad. Los multicubos constituyen una manera más eficiente de almacenar datos dispersos y pueden reducir el efecto de la explosión de los datos debido a los precálculos.

12.5.6 El Futuro de las tecnologías OLAP

El futuro de las tecnologías OLAP va irremediablemente unido a los avances informáticos en multitud de campos. En primer lugar el desarrollo de herramientas OLAP pasa por su adaptación a las modernas tecnologías de comunicaciones entre ordenadores, nos referimos a la conocida red Internet y también a las intranets que puedan crear las empresas (y que cada vez se encuentran más dispuestas a ello ya que son conscientes de la multitud de ventajas que les proporcionan tener sus equipos informáticos unidos y compartidos). Otra tecnología muy importante para OLAP es la tecnología de las bases de datos, refiriéndonos a los avances que en materia de almacenamiento, búsqueda de datos mediante heurísticas y mejora del interfaz están teniendo lugar en el campo de las bases de datos. También consiederemos la relación con el hardware principalmente en materia de rapidez. Por último, y quizás sea el

aspecto más importante al que están ligados los avances en tecnología OLAP, tenemos la cada vez mayor conciencia por parte de empresarios y analistas de la necesidad de una herramienta informática que les ayude en la tarea de analizar los datos y encontrar una relación entre los mismos que explique una determinada situación. (Universidad de Alicante, 1999)

12.6 ÁRBOLES DE DECISIÓN

Si tiene una tabla que contiene datos sobre el comportamiento de los clientes y quiere clasificar estos datos o hacer predicciones, encontrará que las tareas de clasificación y predicción están muy ligadas. Un intento de predecir si un cliente seguro mostrará un tipo de comportamiento seguro, implica una suposición que el cliente corresponde a un acertado tipo de clientes y por lo tanto muestra su tipo de comportamiento. (Adrians, 1996)

12.6.1 DEFINICIÓN

Estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos.

Métodos específicos de árboles de decisión incluyen Arboles de Clasificación y Regresión (CART: Classification And Regression Tree) y Detección de Interacción Automática de Chi Cuadrado (CHAI: Chi Square Automatic Interaction Detection) (Harjinder, 1996)

Los árboles de decisión son una forma de representación sencilla, muy usada entre los sistemas de aprendizaje supervisado, para clasificar ejemplos en un número finito de clases. Se basan en la partición del conjunto de ejemplos según ciertas condiciones que se aplican a los valores de los atributos. Su potencia descriptiva viene limitada por las condiciones o reglas con las que se divide el conjunto de entrenamiento; por ejemplo, estas reglas pueden ser simplemente relaciones de igualdad entre un atributo y un valor, o relaciones de comparación ("mayor que", etc.), etc.

Los sistemas basados en árboles de decisión forman una familia llamada TDIDT (Top-Down Induction of Decision Trees), cuyo representante más conocido es ID3.

ID3 (Interactive Dichotomizer) se basa en la reducción de la entropía media para seleccionar el atributo que genera cada partición (cada nodo del árbol), seleccionando aquél con el que la reducción es máxima. Los nodos del árbol están etiquetados con nombres de atributos, las ramas con los posibles valores del atributo, y las hojas con las diferentes clases. Existen versiones secuenciales de ID3, como ID5R.

C4.5 es una variante de ID3, que permite clasificar ejemplos con atributos que toman valores continuos. ( )

12.7 REGLAS DE ASOCIACIÓN

Las reglas de asociación son siempre definidas sobre atributos binarios, los cuales son usados en una base de datos simple al representar suscripciones a revistas, por lo que, tendríamos que aplanar la tabla antes de ejecutar un algoritmo de asociación. No es muy difícil desarrollar algoritmos que encontrarán ésta asociación en una base de datos grande.

El problema, es que un algoritmo no cubrirá algunas otras asociaciones que son de muy poco valor. No hay muchas mujeres que posean carros deportivos rojos y pequeñas mascotas, así que este es un pequeño subconjunto de los clientes y encontraremos y encontraremos solamente un pequeño subconjunto cuando tenemos

una base de datos grande de clientes nuestra disposición. Tampoco, el número de posibles reglas de asociación que pueden encontrarse en cada base de datos es casi infinita. El problema con las reglas de asociación es que está restringido a encontrar algunas asociaciones que será muy díficil de separar información valuable de ruido, y es necesario introducir alguna medida para distinguir asociaciones interesantes de las no interesantes. Representaremos una regla de asociación en la siguiente forma:

MUSIC_MAG, HOUSE_MAG => CAR_MAG

Esto significa que cualquiera que lee ambas, una revista de música y una de casas es un probable que lea una revista de carros. Recuerde que las reglas de asociación son definidas sobre atributos binarios. Ahora, ¿cuáles asociaciones son interesantes? En primer lugar, buscamos asociaciones que tengan muchos de los ejemplos en la base de datos y termina este proceso con el soporte de una regla de asociación.

Actualmente, las reglas de asociación son sólo útiles en Data Mining si realmente tenemos un bosquejo de que estamos buscando. Esto ilustra que no es un algoritmo que automáticamente nos dará cualquier cosa de interés en la base de datos. Un algoritmo que encuentra muchas reglas probablemente encontrará muchos usos de las reglas, mientras una algoritmo que encuentra sólo un número de asociaciones limitadas, probablemente también perderemos mucha información interesante.

12.8 REDES NEURONALES

Es interesante ver que algunas técnicas de máquinas de aprendizaje son derivadas de paradigmas relacionadas a las diferentes áreas de investigación. Los algoritmos genéticos derivan su inspiración de la biología mientras las redes neuronales son modeladas sobre el cerebro humano.

En la teoría de Freud de psicodinámicas, el cerebro humano fué descrito como una red neuronal, e investigaciones recientes han corroborado este punto. El cerebro humano consiste en un muy grande número de neuronas, cerca de 10^{11} conectadas unas entre otras por sinapsos. Una simple neurona es conectada a otra neurona por una pareja de miles de estos sinapsos. Aunque las neuronas pueden ser descritas como simples bloques construidos del cerebro, el cerebro humano puede realizar tareas muy complejas a pesar de su relativa simplicidad. Esta analogía ofrece un interesante modelo para la creación de más complejas máquinas de aprendizaje, y ha permitido la creación de las llamadas redes neuronales artificiales. (Adrians, 1996)

12.8.1 DEFINICIÓN DE REDES NEURONALES ARTIFICIALES (RNA)

Son modelos predecibles no-lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica. (Harjinder, 1996)

Las RNA están compuestas de un gran número elementos de procesamiento altamente interconectados (Neuronas) trabajando al mismo tiempo para la solución de problemas específicos. Las RNA, tal como las personas, aprenden de la experiencia.

En cualquier caso, se trata de una nueva forma de computación que es capaz de manejar las imprecisiones e incertidumbres que aparecen cuando se trata de resolver problemas relacionados con el mundo real (reconocimiento de formas, toma de decisiones, etc.), ofreciendo soluciones robustas y de fácil implementación.

Están compuestas de muchos elementos sencillos que operan en paralelo, el diseño de la red está determinado mayormente por las conexiones entre sus elementos. Al igual que las conexiones de las neuronas cerebrales.

Han sido entrenadas para la realización de funciones complejas en variados campos de aplicación. Hoy en día pueden ser entrenadas para la solución de problemas que son difíciles para sistemas computacionales comunes o para el ser humano.

El comportamiento complejo se modela conectando un conjunto de neuronas. El aprendizaje o "capacitación" ocurre modificando la "fuerza de conexión" o los parámetros que conectan las capas. Las redes neuronales se acondicionan con muestras adecuadas de la base de datos.

Cada red puede ser construida usando hardware especial pero muchas son sólo programas de software que pueden ser operadas en computadoras normales.

Las características de operación son las siguientes:

PESOS

Las RNA puede tener factores de peso fijos o adaptables. Las que tienen pesos adaptables emplean leyes de aprendizaje para ajustar el valor de la fuerza de un interconección con otras neuronas. Si las neuronas utilizan pesos fijos, entonces su tarea deberá estar previamente definida. Los pesos serán determinados a partir de una descripción completa del problema. Por otra parte, los pesos adaptables son esenciales si no se conoce previamente cual deberá de ser su valor correcto.

DOS TIPOS DE APRENDIZAJE

Existen dos tipos de aprendizaje:

1. **Supervisado:** Ocurre cuando se le proporciona a la red tanto la entrada como la salida correcta, y la red ajusta sus pesos tratando de minimizar el error de su salida calculada. Este tipo de entrenamiento se aplica por ejemplo, en el reconocimiento de patrones.
2. **No supervisado**: Se presenta cuando a la red se le proporcionan únicamente los estímulos, y la red ajusta sus interconecciones basándose únicamente es sus estímulos y la salida de la propia red. Las leyes de aprendizaje determinan como la red ajustará sus pesos utilizando una función de error o algún otro criterio. La ley de aprendizaje adecuada se determina en base a la naturaleza del problema que se intenta resolver.

Las RNA adaptables tienen **dos fases en su operación:**

1. Entrenamiento de la red: El usuario proporciona a la red un número "adecuado" de estímulos de entrada, y de salida, la red entonces ajusta su pesos de interconección o sinápsis hasta que la salida de la red esta "lo suficientemente cerca" de la salida correcta.

2. **Recuperación de lo aprendido:** A la red se le presenta un conjunto de estímulos de entrada y esta simplemente calcula su salida. Cuando la red emplea entrenamiento no supervisado, algunas veces será necesario que reajuste su sinápsis durante la fase de recuperación.

La gran **diferencia** del empleo de las redes neuronales **en relación con otras aplicaciones** de la computación radica en que **no son algorítmicas**, esto es no se programan haciéndoles seguir una secuencia predefinida de instrucciones. Las RNA generan ellas mismas sus propias "reglas", para asociar la respuesta a su entrada; es decir, aprende por ejemplos y de sus propios errores.

El conocimiento de una RNA se encuentra en la función de activación utilizada y en los valores de sus pesos.

12.8.2 ASOCIAR Y GENERALIZAR SIN REGLAS COMO EN EL CEREBRO HUMANO

Las redes neuronales formadas por los perceptrones se interconectan en forma muy similar a como las neuronas humanas se disponen en la corteza cerebral humana, y lo más importante, son capaces de asociar y generalizar sin reglas. Han sido utilizadas con gran éxito para reconocer retornos de sonar bajo el agua, escritura a mano, voz, topografía de terrenos, controlar brazos de robots, evaluar datos personales, modelar fenómenos cognocitivos, y, predecir tendencias financieras.

REQUIEREN DE ALGÚN TIPO DE PATRÓN.

La clase de problemas que mejor se resuelven con las redes neuronales son los mismos que el ser humano resuelve mejor: Asociación, evaluación, y reconocimiento de patrones. Las redes neuronales son perfectas para problemas que son muy difíciles de calcular pero que no requieren de respuestas perfectas, sólo respuestas rápidas y buenas. Tal y como acontece con el escenario bursátil en el que se quiere saber ¿compro?, ¿vendo?, ¿mantengo?, o en el reconocimiento cuando se desea saber ¿se parece? ¿es el mismo pero tienen una ligera modificación?

Por otra parte, las redes neuronales son muy malas para cálculos precisos, procesamiento serie, y no son capaces de reconocer nada que no tenga inherentemente algún tipo de patrón. Es por esto, que no pueden predecir la lotería, ya por definición es un proceso al azar.

12.8.3 TIPOS DE REDES NEURONALES

Existen varias formas de hacer las conexiones en una RNA, así como existen varias formas de conectar neuronas biológicas en el cerebro. Cada tipo sirve para diferentes procesos, el elegir la correcta topología y sus características, es imprescindible para lograr fácilmente la solución del problema.

A continuación analizaremos algunas topologías de RNA:

1. Perceptron
2. Backpropagation
3. Hopfield
4. Kohonen

12.8.3.1 REDES PERCEPTRON

ANTECEDENTES

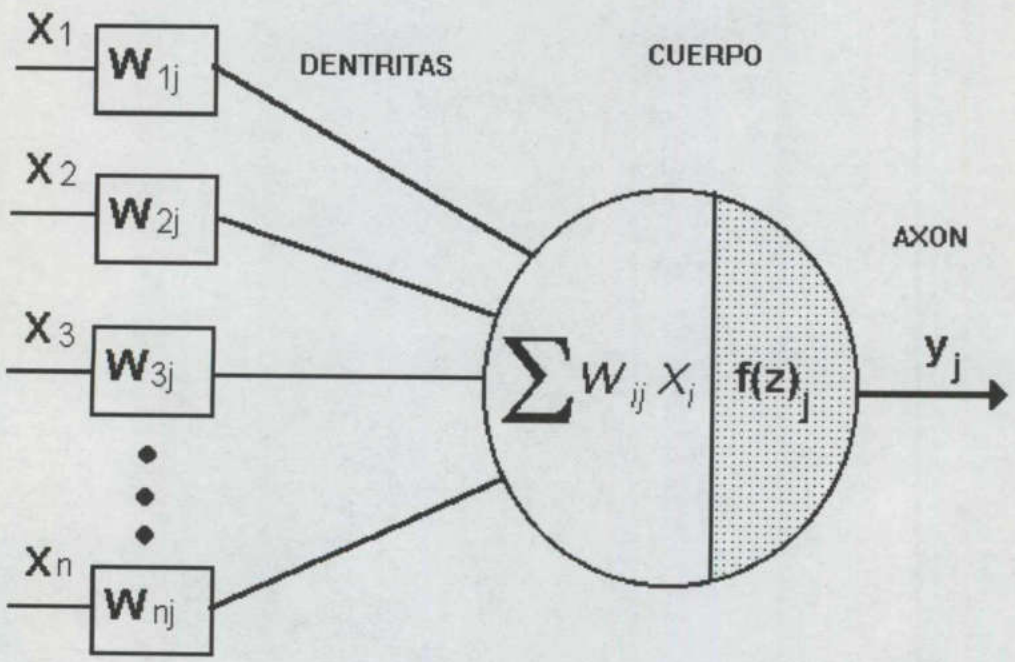
En 1943, Warren McCulloc y Walter Pitts originaron el primer modelo de operación neuronal, el cual fué mejorado en sus aspectos biológicos por Donald Hebb en 1948. En 1962 Bernard Widrow propuso la regla de aprendizaje Widrow-Hoff, y Frank Rosenblatt desarrolló una prueba de convergencia, y definió el rango de problemas para los que su algoritmo aseguraba una solución. El propuso los 'Perceptrons' como herramienta computacional.

FUNCIONAMIENTO

En la siguiente figura se representa una neurona "artificial", que intenta modelar el comportamiento de la neurona biológica. Aquí el cuerpo de la neurona se representa como un sumador lineal de los estímulos externos zj, seguida de una función no lineal yj = f(zj). La función f(zj) es llamada la función de activación, y es la función que utiliza la suma de estímulos para determinar la actividad de salida de la neurona.

Este modelo se conoce como perceptrón de McCulloch-Pitts, y es la base de la mayor parte de las arquitectura de las RNA que se interconectan entre sí. Las neuronas emplean funciones de activación diferentes según la aplicación, algunas veces son funciones lineales, otras funciones sigmoidales (p.ej. la tanh), y otras funciones de umbral de disparo. La eficiencia sináptica se representa por factores de peso de interconección wij, desde la neurona i, hasta la neurona j.

Los pesos pueden ser positivos (excitación) o negativos (inhibición). Los pesos junto con las funciones f(z) dictan la operación de la red neuronal. Normalmente las funciones no se modifican de tal forma que el estado de la red neuronal depende del valor de los factores de peso (sinápsis) que se aplica a los estímulos de la neurona.



En un perceptrón, cada entrada es multiplicada por el peso W correspondiente, y los resultados son sumados, siendo evaluados contra el valor de umbral, si el resultado es mayor al mismo, el perceptrón se activa.

LIMITANTES

El perceptrón es capaz tan sólo de resolver funciones definidas por un hiperplano (objeto de dimensión N-1 contenida en un espacio de dimensión N). que corte un espacio de dimensión N. Un ejemplo de una función que no puede ser resuelta es el operador lógico XOR.

Una explicación mas sencilla de un hiperplano sería, hablando en un plano de dos dimensiones, una línea que separa a los elementos existentes en dos grupos. El perceptrón sólo puede resolver una función, si todos los posibles resultados del problema pueden separarse de ésta forma (en dos secciones) es decir, que no se combinen entre sí.

ENTRENAMIENTO

El entrenamiento de un perceptrón es por medio de la regla de aprendizaje delta: Para cada peso W se realiza un ajuste dW según la regla:

$dW = LR (T - Y) X$

Donde LR es la razón de aprendizaje, T el valor deseado, Y el valor obtenido, y X la entrada aplicada al perceptrón.

12.8.3.1.1 TIPOS DE PERCEPTRÓN

El Perceptrón básico de dos capas (entrada con neuronas lineales, analógicas, y la de salida con función de activación de tipo escalón, digital) solo pude establecer dos regiones separadas por una frontera lineal en el espacio de patrones de entrada, donde se tendría un hiperplano.

Un Perceptrón con tres niveles de neuronas puede formar cualquier región convexa en este espacio. Las regiones convexas se forman mediante la intelección entre las regiones formadas por cada neurona de la segunda capa, cada uno de estos elementos

se comporta como un Perceptrón simple, activándose su salida para los patrones de un lado del hiperplano.

Un Perceptrón con cuatro capas puede generar regiones de decisión arbitrariamente complejas. El proceso de separación en clases que se lleva a cabo consiste en la partición de la región deseada en pequeños hipercubos. Cada hipercubo requiere 2n neuronas en la segunda capa (siendo n el numero de entradas a la red), una por cada lado del hipercubo, y otra en la tercera capa, que lleva a cabo el and lógico de la salida de los nodos del nivel anterior. La salida de los nodos de este tercer nivel se activaran solo para las entradas de cada hipercubo. Los hipercubos se asignan a la región de decisión adecuada mediante la conexión de la salida de cada nodo del tercer nivel solo con la neurona de salida (cuarta capa) correspondiente a la región de decisión en la que este comprendido el hipercubo llevándose a cabo una operación lógica Or en cada nodo de salida. Este procedimiento se pude generalizar de manera que la forma de las regiones convexas sea arbitraria, en lugar de hipercubos.

En teoría, el Perceptrón de 4 capas puede resuelve una gran variedad de problemas cuyas entradas sean analógicas, la salida sea digital y sea linealmente separable. El problema práctico radica en el numero de neuronas, en el numero idóneo de capas ocultas, la extensión de la función de activación, el tiempo de entrenamiento de la red, las implicaciones en la generación de ruido (al tener un numero excesivo de neuronas) en contraparte con la ventaja de tener un sistema tolerante a fallas al tener un numero de neuronas redundante.

12.8.3.1.2 APLICACIONES DEL PERCEPTRÓN.

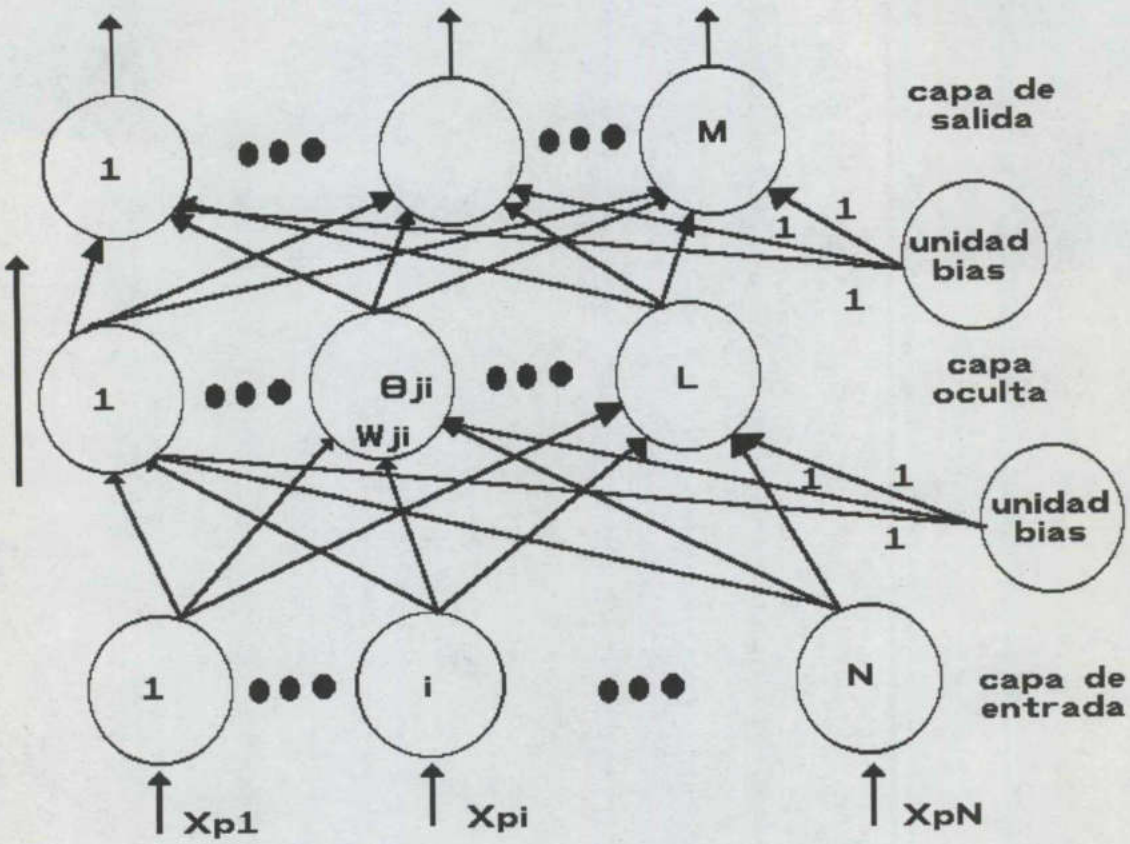
El rango de tareas que el Perceptrón puede manejar es mucho mayor que simples decisiones y reconocimiento de patrones. Por ejemplo, se puede entrenar una red para formar el tiempo pasado de los verbos en ingles, leer texto en ingles y manuscrito. El Perceptrón multicapa (MLP) puede ser usado para la predicción de una serie de datos en el tiempo; tal ha sido su éxito en la medición de la demanda de gas y electricidad, además de la predicción de cambios en el valor de los instrumentos financieros.

Predicción de mercados financieros, diagnósticos médicos, el Perceptrón como una red codificadora, el Perceptrón aprende a sumar enteros.

NETtalk es un Perceptrón que es capaz de transformar texto en ingles en sonido individual (representaciones fonéticas) y la pronunciación con la utilización de un sintetizador de voz; cuenta con aproximadamente 300 nodos de neuronas (siendo 80 en la capa escondida) y 20,000 conexiones individuales.

El perceptrón solo es el ejemplo más elemental de una red neuronal, de hecho, no puede siquiera ser considerado una "red", puesto que no intervienen otros elementos. Si se combinan varios perceptrones en una "capa", y los estímulos de entrada después se suman tendremos ya una red neuronal. Una red neuronal muy eficaz para resolver

fundamentalmente problemas de reconocimiento de patrones es la red neuronal de propagación hacia atrás, en inglés back propagation network.



En esta red, se interconectan varias unidades de procesamiento en capas, las neuronas de cada capa no se interconectan entre sí. Sin embargo, cada neurona de una capa proporciona una entrada a cada una de las neuronas de la siguiente capa, esto es, cada neurona transmitirá su señal de salida a cada neurona de la capa siguiente.

12.8.3.2 HOPFIELD

Las redes de Hopfield son redes de adaptación probabilística, recurrentes, funcionalmente entrarían en la categoría de las memorias autoasociativas, es decir, que aprenden a reconstruir los patrones de entrada que memorizaron durante el entrenamiento. Son arquitecturas de una capa con interconexión total, funciones de activación booleana de umbral (cada unidad puede tomar dos estados, 0 o 1, dependiendo de si la estimulación total recibida supera determinado umbral), adaptación probabilística de la activación de las unidades, conexiones recurrentes y simétricas, y regla de aprendizaje no supervisado. Mientras que las redes en cascada (no recurrentes) dan soluciones estables, los modelos recurrentes dan soluciones inestables (dinámicas), lo que no siempre es aconsejable.

La principal aportación de Hopfield consistió precisamente en conseguir que tales redes recurrentes fueran así mismo estables. Imaginó un sistema físico capaz de operar como una memoria autoasociativa, que almacenara información y fuera capaz de recuperarla aunque la misma se hubiera deteriorado.

La Red de Hopfield es recurrente y completamente conectada. Funciona como una memoria asociativa no lineal que puede almecenar internamente patrones presentados de forma incompleta o con ruido. De esta forma puede ser usada como una herramienta de optimización. El estado de cada neurona puede ser actualizado un

número indefinido de veces, independientemente del resto de las neuronas de la red pero en paralelo.

12.8.3.3 BOLTZMANN

En la Máquina de Boltzmann, generalización de la red de Hopfield que incluye unidades ocultas, la operación de actualización se basa en un concepto de termodinámica estadística conocido como "simulated annealing". La red de Hopfield, la máquina de Boltzmann y un derivado conocido como la máquina del teorema de campo medio se han utilizado en aplicaciones de segmentación y restauración de imágenes y optimización combinacional.

12.8.3.3.1 CARACTERÍSTICAS

La red de Hopfield consiste en un conjunto de N elementos de procesado interconectadas que actualizan sus valores de activación de forma asincrona e

independiente del resto de las elementos de procesado. Todos los elementos son a la vez de entrada y salida. Los valores de activación son binarios.

12.8.3.3.2 FUNCIONAMIENTO

A cada estado de la red se le puede atribuir una cierta cantidad de energía, el sistema evoluciona tratando de disminuir la energía mediante un proceso de relajación, hasta alcanzar un mínimo (valle) donde se estabiliza. Los mínimos de energía se corresponden con los recuerdos almacenados durante el aprendizaje de la red.

Ante la presentación de un estímulo nuevo se obtendrá una configuración inicial más o menos parecida a alguno de los estímulos almacenados, el sistema evolucionará hasta caer en una configuración estable que representa el recuerdo asociado a ese estímulo. Si la configuración inicial discrepa mucho de los recuerdos almacenados podemos alcanzar algún mínimo que no se corresponde a ningún recuerdo almacenado, recuperando en ese caso una información espuria, o podríamos no alcanzar ningún mínimo, quedando inestable: en ese caso diríamos que la red está "confundida", no es capaz de reconocer el estímulo, no recuerda.

Una tercera posibilidad es que al cabo de unos pasos de evolución empiece a repetir periódicamente una secuencia definida de estados; con esta dinámica se han

modelado ciertas excitaciones nerviosas que regulan acciones rítmicas y repetitivas; y se ha tratado de reproducir la memoria de secuencias temporales, por ejemplo, el recuerdo de melodías.

Las redes de Hopfield se han aplicado a campos como la percepción, el reconocimiento de imágenes y optimización de problemas, mostrando gran inmunidad al ruido y robustez. Incluso se han llegado a desarrollar chips específicos para este tipo redes. Hopfield ha mostrado como aplicar los mismos principios con funciones de activación continuas como la función sigmoidal, con muy pocas modificaciones.

Pero pese a sus evidentes ventajas no están exentas de problemas:

- El número máximo de patrones no correlacionados que puede almacenar es igual al 15% del número de neuronas de la red.
- Requieren mucho tiempo de procesamiento hasta converger a una -solución estable, lo que limita su aplicabilidad.
- Tendencia a caer en mínimos locales, como en las redes de retropropagación. La solución pasa por aplicar los métodos estadísticos que ya comentamos en el apartado dedicado a las redes de retropropagación, el equilibrio termodinámico simulado.

12.8.3.4 KOHONEN.

Existen evidencias que demuestran que en el cerebro existen neuronas que se organizan en muchas zonas, de forma que las informaciones captadas del entorno a través de los órganos sensoriales se representan internamente en forma de capas bidimensionales. Por ejemplo, en el sistema visual se han detectado mapas del espacio visual en zonas de córtex (capa externa del cerebro). También en el sistema auditivo se detecta organización según la frecuencia a la que cada neurona alcanza la mayor respuesta (organización tonotópica).

Aunque en gran medida esta organización neuronal está predeterminada genéticamente, es probable que de ella se origine medienta el aprendizaje. Esto sugiere, por tanto, que el cerebro podría poseer la capacidad inherente de formar mapas topológicos de las informaciones recibidas del exterior. De hecho, esta teoría podría explicar su poder de operar con elementos semánticos: algunas areas del cerebro simplemente podrían crear y ordenar neuronas especializadas o grupos con caracteristicas de alto nivel y sus combinaciones. Se trataría, en definitiva, de construir mapas espaciales para atributos y características.

12.8.3.4.1 HISTORIA

A partir de estas ideas, T. Kohonen presentó en 1982 un sistema con un comportamiento semejante. Se trataba de un modelo de red neuronal con capacidad para formar mapas de características de manera similar a como ocurre en el cerebro.
El objetivo de Kohonen era demostrar que en un estímulo externo (información de entrada) por si solo, suponiendo una estructura propia y una descripción funcional del comportamiento de la red, era suficiente para forzar la formacion de mapas.
Este modelo tiene dos variantes, denominadas LVQ (Learning Vector Quantization) y TPM (Topology-Preserving Map) o SOM (Self-Organizating Map). Ambas se basan en el principio de formación de mapas topológicos para establecer características comunes entre las informaciones (vectores) de entrada a la red, aunque difieren en las dimensiones de éstos, siendo de una soladimensión en el caso de LVQ, y bidimensional, e incluso tridimensional, en la red TPM.

12.8.3.4.3 CARACTERÍSTICAS

Pertenece a la categoría de las redes competitivas o mapas de autoorganización, es decir, aprendizaje no supervisado. Poseen una arquitectura de dos capas (entrada-

salida) (una sola capa de conexiones), funciones de activación lineales y flujo de información unidireccional (son redes en cascada).

Las unidades de entrada reciben datos continuos normalizados, se normalizan así mismo los pesos de las conexiones con la capa de salida. Tras el aprendizaje de la red, cada patrón de entrada activará una única unidad de salida.

El objetivo de este tipo de redes es clasificar los patrones de entrada en grupos de características similares, de manera que cada grupo activará siempre la(s) misma(s) salida(s). Cada grupo de entradas queda representado en los pesos de las conexiones de la unidad de salida triunfante. La unidad de salida ganadora para cada grupo de entradas no se conoce previamente, es necesario averiguarlo después de entrenar a la red.

12.8.3.4.3 ARQUITECTURA

En la arquitectura de la versión original (LVQ) del modelo Kohonen no existen conexioines hacia atrás. Se trata de una de las N neuronas entrada y M de salida. Cada una de las N neuronas de entrada se conecta a las M de salida a través de conexiones hacia adelante (feedfoward).

Entre las neuronas de la capa de salida, puede decirse que existen conexiones laterales de inhibición (peso negativo) implícitas, pues aunque no estén conectadas, cada una de las neuronas va a tener cierta influencia sobre sus vecinas. El valor que se asigne a los pesos de las conexiones hacia adelante entre las capas de entrada y salida (Wji) durante el proceso de aprendizaje de la red va a depender precisamente de esta interacción lateral.

La influencia que una neurona ejerce sobre las demás es función de la distancia entre ellas, siendo muy pequeñas cuando estan muy alejadas. Es frecuente que dicha influencia tenga la forma de un sombrero mexicano.

Por otra parte, la versión del modelo denominada TPM (Topology Preserving Map) trata de establecer una correspondencia entre los datos de entrada y un espacio bidimensional de salida, crenado mapas topológicos de dos dimensiones, de tal forma que ante datos de entrada con características comunes se deben activar neuronas situadas en próximas zonas de la capa de salida.

12.8.3.4.4 APRENDIZAJE

Supongamos que tenemos patrones de entrada n-dimensionales.

1. Aleatorizar los pesos de las conexiones. Normalizar los pesos de las conexiones incidentes de cada unidad de salida sobre la unidad: dividir cada conexión por la raíz cuadrada de la suma de los cuadrados de las conexiones de cada unidad. Normalizar igualmente los datos de entrada
2. Aplicar un patrón de entrada.
3. Calcular alguna medida de similitud/disimilitud (producto interno, distancia euclídea o de Mahalanobis, etc.) entre las entradas y los pesos de las conexiones.
4. La unidad de salida con los pesos más parecidos al patrón de entrada es declarada ganadora. El vector de pesos de la unidad ganadora, se convierte en el centro de un grupo de vectores cercanos a él.
5. Modificar los pesos de los vectores de pesos Wj "cercanos" a Wc (distancia menor a D).

   De esta manera conseguimos que los vectores de pesos de la unidad ganadora y de su "vecindario" se parezcan cada vez más al patrón de entrada que hace ganar a esa unidad.
6. Repetir los pasos 1 a 4 con todos los patrones de entrada.

A medida que avanza el aprendizaje hay que ir reduciendo D y a. Kohonen recomienda empezar con un valor de a cercano a 1 y reducirlo gradualmente hasta 0.1. D puede empezar valiendo la máxima distancia existente entre los pesos de las conexiones al principio y acabar siendo tan pequeño que no quede ninguna unidad en el vecindario de la unidad ganadora. En ese momento solo se entrenará una unidad, que al final tendrá su vector de pesos igual al vector de entrada.

La precisión de la clasificación de los patrones de entrada aumenta con el número de ciclos de aprendizaje. Kohonen recomienda una cantidad de ciclos no inferior a 500 veces el número de neuronas de salida para obtener buenos resultados.

12.8.3.4.5 APLICACIÓN

Una vez entrenada, podemos usar a la red para clasificar patrones de entrada similares en el espacio n-dimensional. Una clase o grupo de patrones similares tiende a controlar una neurona específica, que representará el centro de una esfera n-dimensional (de radio unitario, pues normalizamos los datos sobre la unidad). Esa neurona resultará la más activada frente a los patrones más parecidos a su vector de pesos.

Después del aprendizaje, la clasificación consiste en presentar una entrada y seleccionar la unidad más activada. Además, el vector de pesos nos servirá para reconstruir el patrón de entrada. (TREC internet, 1999)

12.9 ALGORITMOS GENÉTICOS

12.9.1 DEFINICIÓN

- Técnicas de optimización que usan procesos tales como combinaciones genéticas, mutaciones y selección natural en un diseño basado en los conceptos de evolución. (Harjinder, 1996)

- Los Algoritmos Genéticos (AGs) son métodos adaptativos que pueden usarse para resolver problemas de búsqueda y optimización. Están basados en el proceso genético de los organismos vivos. A lo largo de las generaciones, las poblaciones evolucionan en la naturaleza de acorde con los principios de la selección natural y la supervivencia de los más fuertes, postulados por Darwin. Por imitación de este proceso, los

Algoritmos Genéticos son capaces de ir creando soluciones para problemas del mundo real. La evolución de dichas soluciones hacia valores óptimos del problema depende en buena medida de una adecuada codificación de las mismas.

Los principios básicos de los Algoritmos Genéticos fueron establecidos por Holland

Los Algoritmos Genéticos usan una analogía directa con el comportamiento natural. Trabajan con una población de individuos, cada uno de los cuales representa una solución factible a un problema dado. A cada individuo se le asigna un valor ó puntuación, relacionado con la bondad de dicha solución. En la naturaleza esto equivaldría al grado de efectividad de un organismo para competir por unos determinados recursos. Cuanto mayor sea la adaptación de un individuo al problema, mayor será la probabilidad de que el mismo sea seleccionado para reproducirse, cruzando su material genético con otro individuo seleccionado de igual forma. Este cruce producirá nuevos individuos – descendientes de los anteriores – los cuales comparten algunas de las características de sus padres. Cuanto menor sea la adaptación de un individuo, menor será la probabilidad de que dicho individuo sea seleccionado para la reproducción, y por tanto de que su material genético se propague en sucesivas generaciones.

De esta manera se produce una nueva población de posibles soluciones, la cual reemplaza a la anterior y verifica la interesante propiedad de que contiene una mayor

proporción de buenas características en comparación con la población anterior. Así a lo largo de las generaciones las buenas características se propagan a través de la población. Favoreciendo el cruce de los individuos mejor adaptados, van siendo exploradas las áreas más prometedoras del espacio de búsqueda. Si el Algoritmo Genético ha sido bien diseñado, la, población convergerá hacia una solución óptima. del problema.

El poder de los Algoritmos Genéticos proviene del hecho de que se trata de una técnica robusta, y pueden tratar con éxito una gran variedad de problemas provenientes de diferentes áreas, incluyendo aquellos en los que otros métodos encuentran dificultades. Si bien no se garantiza que el Algoritmo Genético encuentre la solución óptima, del problema, existe evidencia empírica de que se encuentran soluciones de un nivel aceptable, en un tiempo competitivo con el resto de algoritmos de optimización combinatoria. En el caso de que existan técnicas especializadas para resolver un determinado problema, lo más probable es que superen al Algoritmo Genético, tanto en rapidez como en eficacia. El gran campo de aplicación de los Algoritmos Genéticos se relaciona con aquellos problemas para los cuales no existen técnicas especializadas. Incluso en el caso en que dichas técnicas existan, y funcionen bien, pueden efectuarse mejoras de las mismas hibridándlas con los Algoritmos Genéticos.

12.9.2 EL ALGORITMO GENÉTICO SIMPLE

El Algoritmo Genético Simple, también denominado Canónico, se representa en la siguiente figura. Como se verá a continuación, se necesita una codificación o representación del problema, que resulte adecuada al mismo. Además se requiere una función de ajuste ó adaptación al problema, la cual asigna un número real a cada posible solución codificada. Durante la ejecución del algoritmo, los padres deben ser seleccionados para la reproducción, a continuación dichos padres seleccionados se cruzarán generando dos hijos, sobre cada uno de los cuales actuará un operador de mutación. El resultado de la combinación de las anteriores funciones será un conjunto de individuos (posibles soluciones al problema), los cuales en la evolución del Algoritmo Genético formarán parte de la siguiente población.

```
BEGIN /* Algoritmo Genetico Simple */
    Generar una poblacion inicial.
    Computar la funcion de evaluacion de cada individuo.
    WHILE NOT Terminado DO
    BEGIN /* Producir nueva generacion */
        FOR Tamaño poblacion/2 DO
        BEGIN /*Ciclo Reproductivo */
            Seleccionar dos individuos de la anterior generacion,
            para el cruce (probabilidad de seleccion proporcional
            a la funcion de evaluacion del individuo).
            Cruzar con cierta probabilidad los dos
            individuos obteniendo dos descendientes.
            Mutar los dos descendientes con cierta probabilidad.
            Computar la funcion de evaluacion de los dos
            descendientes mutados.
            Insertar los dos descendientes mutados en la nueva generacion.
        END
        IF la poblacion ha convergido THEN
            Terminado := TRUE
    END
END
```

Figura 1: Pseudocódigo del Algoritmo Genético Simple

12.9.3 CODIFICACIÓN.

Se supone que los individuos (posibles soluciones del problema), pueden representarse como un conjunto de parámetros (que denominaremos penes), los cuales agrupados forman una ristra de valores (a menudo referida como cromosoma). Si bien el alfabeto utilizado para representar los individuos no debe necesariamente estar constituido por el {0, l), buena parte de la teoría en la que se fundamentan los Algoritmos Genéticos utiliza dicho alfabeto. En términos biológicos, el conjunto de parámetros representando un cromosoma particular se denomina fenotipo. El fenotipo

contiene la información requerida para construir un organismo, el cual se refiere como genotipo. Los .mismos términos se utilizan en el campo de los Algoritmos Genéticos. La adaptación al problema de un individuo depende de la evaluación del genotipo. Esta última puede inferirse a partir del fenotipo, es decir puede ser computada a partir del cromosoma, usando la función de evaluación. La función de adaptación debe ser diseñada parra cada problema de manera específica. Dado un cromosoma particular, la función de adaptación le asigna un número real, que se supone refleja el nivel de adaptación al problema del individuo representado por el cromosoma.

Durante la fase reproductiva se seleccionan los individuos de la población para cruzarse y producir descendientes, que constituirán, una vez. mutados, la siguiente generación de individuos. La selección de padres se efectúa al azar usando un procedimiento que favorezca a los individuos mejor adaptados, ya que a cada individuo se le asigna una probabilidad de ser seleccionado que es proporcional a su función de adaptación. Este procedimiento se dice que está basado en la ruleta sesgada. Según dicho esquema, los individuos bien adaptados se escogerán probablemente varias veces por generación, mientras que, los pobremente adaptados al problema, no se escogerán más que de vez en cuando.

Una vez seleccionados dos padres, sus cromosomas se combinan, utilizando habitualmente los operadores de cruce y mutación. Las formas básicas de dichos operadores se describen a continuación.

El *operador de cruce*, coge dos padres seleccionados y corta sus ristras de cromosomas en una posición escogida al azar, para producir dos subristras iniciales y dos subristras finales. Después se intercambian las subristras finales, produciéndose dos nuevos cromo-somas completos (véase la Figura). Ambos descendientes heredan genes de cada uno de los padres. Este operador se conoce como operador de cruce basado en un punto. Habitualmente el operador de cruce no se aplica a todos los pares de individuos que han sido seleccionados para emparejarse, sino que se aplica de manera aleatoria, normalmente con una probabilidad comprendida entre 0.5 y 1.0. En el caso en que el operador de cruce no se aplique, la descendencia se obtiene simplemente duplicando los padres.

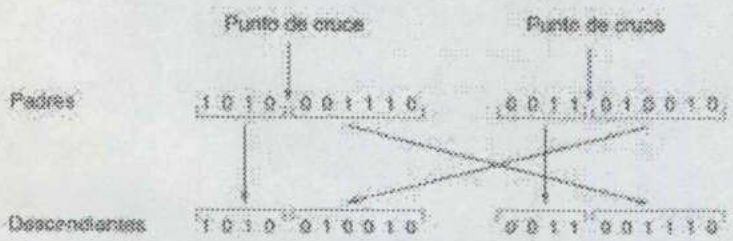


Figura 2: Operador de cruce basado en un punto

El operador de mutación se aplica a cada hijo de manera individual, y consiste en la alteración aleatoria (normalmente con probabilidad pequeña) de cada gen componente del cromosoma. La Figura de abajo muestra la mutación del quinto gen

del cromosoma. Si bien puede en principio pensarse que el operador de cruce es más importante que el operador de mutación, ya que proporciona una exploración rápida del espacio de búsqueda, éste último asegura que ningún punto del espacio de búsqueda tenga probabilidad cero de ser examinado, y es de capital importancia para asegurar la convergencia de los Algoritmos Genéticos.

Figura 3: Operador de mutación

Si el Algoritmo Genético ya está correctamente implementado, la población evolucionará a lo largo de las generaciones sucesivas de tal manera que la adaptación media extendida a todos los individuos de la población, así como la adaptación del mejor individuo se irán incrementando hacia el óptimo global. El concepto de convergencia está relacionado con la progresión hacia la uniformidad: un gen ha convergido cuando al menos el 95 % de los individuos de la población comparten el mismo valor para dicho gen. Se dice que la población converge cuando todos los genes han convergido. Se puede generalizar dicha definición al caso en que al menos un poco de los individuos de la población hayan convergido.

La Figura muestra como varía la adaptación media y la mejor adaptación en un Algoritmo Genético Simple típico.

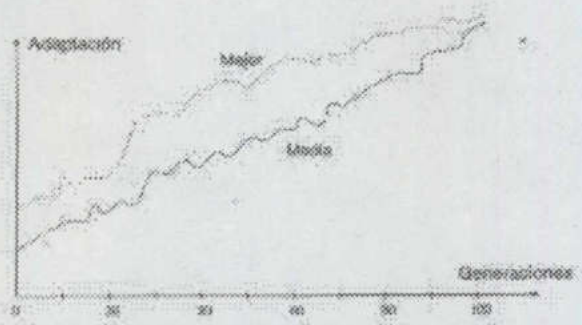


Figura 4: Adaptación media y mejor adaptación en un Algoritmo Genético Simple

A medida que el número de generaciones aumenta, es más probable que la adaptación media se aproxime a la del mejor individuo.

Evaluación del Comportamiento de los Algoritmos Genéticos.

Las tres medidas de evaluación, fueron introducidas por De Jong, y se conocen como:

- evaluación on-line
- evaluación og-line
- evaluación basada en el mejor (Larrañaga, 2000)

La fórmula para la construcción de un algoritmo genético para la solución de un problema es la siguiente:

1. Idear un buen y elegante código del problema en términos de cadenas de un alfabeto limitado.

2. Inventar un ambiente artificial en la computadora donde las soluciones puedan unir una lucha con otra. Proveer un rango objetivo de juicio éxitoso o fallado, términos profesionales llamados función de capacidad.

3. Desarrollar una forma las cuales puedan ser posibles soluciones combinadas. Estas son llamadas operaciones cruzadas, en las cuales las cadenas madre y padre son cortas y después de cambiarse, se pegan. En la reproducción, los tipos de operaciones de mutación pueden ser aplicadas.

4. Proveer una buena variedad de población inicial y hacer que la computadora realice "evolución", removiendo las soluciones malas de cada generación y reemplazarlo con prole o mutaciones de buenas soluciones. Para cuando una familia de soluciones exitosas han sido producidas.

Los algoritmos genéticos pueden ser vistos como un tipo de estrategia de meta aprendizaje. En cinco años se ha visto el desarrollo de algunas aproximaciones híbridas, en las cuales las redes neuronales han sido usadas para crear entradas para algoritmos genéticos o alternativamente algoritmos genéticos para optimizar la salida de redes neuronales. Actualmente, la programación genética es ampliamente usada en mercados financieros y para aplicaciones seguras. (Adrians, 1996)

12.10 MÉTODO DEL VECINO MÁS CERCANO

Una técnica que clasifica cada registro en un conjunto de datos basado en una combinación de las clases de los *k* registros más similares a él en un conjunto de datos históricos (donde *k* □ 1). Algunas veces se llama la técnica del vecino *k*-más cercano.

12.11 REGLA DE INDUCCIÓN

La extracción de reglas if-then de datos basados en significado estadístico.

Muchas de estas tecnologías han estado en uso por más de una década en herramientas de análisis especializadas que trabajan con volúmenes de datos relativamente pequeños. Estas capacidades están ahora evolucionando para integrarse directamente con herramientas OLAP y de Data Warehousing.

Pueden emplearse diferentes criterios para clasificar los sistemas de minería de datos y, en general, los sistemas de aprendizaje inductivo en ordenadores:

Dependiendo del objetivo para el que se realiza el, pueden distinguirse sistemas para:

- clasificación: clasificar datos en clases predefinidas.
- Regresión: función que convierte datos en valores de una función de predicción.
- agrupamiento de conceptos: búsqueda de conjuntos en los que agrupar los datos.
- Compactación: búsqueda de descripciones más compactas de los datos.
- modelado de dependencias: dependencias entre las variables de los datos.
- detección de desviaciones: búsqueda de desviaciones importantes de los datos respecto de valores anteriores o medios, etc.

Dependiendo de la tendencia con que se aborde el problema, se pueden distinguir tres grandes líneas de investigación o paradigmas:

1. **Sistemas conexionistas**: redes neuronales
2. **Sistemas evolucionistas**: algoritmos genéticos
3. **Sistemas simbólicos.**

Dependiendo del lenguaje utilizado para representar el conocimiento, se pueden distinguir:

- Representaciones basadas en la lógica de proposiciones.
- Representaciones basadas en lógica de predicados de primer orden.
- Representaciones estructuradas.
- Representaciones a través de ejemplos y
- Representaciones no simbólicas como las redes neuronales. (Gómez, 1998)

12.12 VENDEDORES Y FABRICANTES DE DATA MINING

Existen docenas de vendedores de Data Mining, aunque algunas industrias se han comenzado a consolidar, por ahora, no es un mercado limpio. y muchos de los productos son caros y complejos de usar.

Desarrollados comúnmente para workstations bajo Unix, para matemáticos y estadísticos y no especialmente para típicas base de datos.

El análisis de mercado de las herramientas de Data Mining de Herb Edelstein, disponible en www.twocrows.com es simplemente el mejor recurso de información acerca del mercado de Data Mining.

Edelstein provee un análisis de los siguientes vendedores y sus herramientas:

- AbTech Software (ModelQuest MarketMiner)
- *Angoss Software (KnowledgeSEEKER, KnowledgeSTUDIO)
- Attar Software (XpertRule Miner)
- Business Objects (BusinessMiner)
- Cognos Software (4Thought, Scenario)
- Group 1 (Model 1)

- HNC Software Inc. (DataBase Mining Marksman)
- Integral Solutions (Clementine, acquired by SPSS in 1998)
- IBM (Intelligent Miner)
- Magnify (PATTERN)
- MathSoft (S-Plus)
- NCR (TeraMiner)
- NeoVista Software (Decision Series)
- Quadstone (Decisionhouse)
- Salford Systems (CART, MARS)
- *SAS Institute (Enterprise Miner)
- *Silicon Graphics (MineSet)
- *SPSS (Base, AnswerTree, Neural Connection)
- Tandem Division of Compaq
- Thinking Machines (Darwin, acquired by Oracle in 1999)
- Torrent Systems (Orchestrate Analytics)
- Trajecta (dbProphet)
- Unica Technologies (PRW)
- Urban Science Applications (GainSmarts)

* Estos vendedores han colaborado con Microsoft para crear el OLE DB.

CONCLUSIONES

Sin duda alguna, la técnica de Data Mining es una de las herramientas que ha alcanzado un auge impresionante dentro de las grandes organizaciones que están dispuestas a realizar todo un proceso de reingeniería para optimizar el uso y explotación de su información. Grandes bases de datos son filtradas con la finalidad de encontrar soluciones óptimas a problemas a los que se enfrenta y, no sólo eso, también promueve soluciones proactivas para que, en un futuro se tomen decisiones con mayor certidumbre.

Estas herramientas son realmente interesantes cuando a través de disciplinas como la inteligencia artificial, las redes neuronales, los árboles de decisión y otras tecnologías como lo son las fabulosas herramientas OLAP logran hacer del Data Mining una tecnología basada en el aprendizaje automatizado. Actualmente, existen infinidad de herramientas de Data Mining, desarrollados por fabricantes de renombre tales como Microsoft, IBM, ORACLE, etc., que han hecho posible que el software de Data Mining sea cada día más robusto respecto a las soluciones que ésta ofrece y con interfaces gráficas de alta calidad que permiten una visualización de los resultados de forma clara y concisa.

Una solución que podríamos considerar como apropiada para "minar" datos críticos y eliminar los obsoletos es, sin duda alguna, la aplicación de las técnicas de Data Mining, considerada entre los Sistemas de Soporte a la Decisión y los Sistemas de Soporte para Ejecutivos como herramienta imprescindible para la correcta toma de decisiones. Se espera que en los próximos años, Data Mining así como Data Warehouse y Datamarts sean las técnicas más usadas en la manipulación de enormes bases de datos. De ahí su importancia de estudio y desarrollo.

GLOSARIO DE TÉRMINOS

Pivoting (Girar) . Es la habilidad de cambiar la jerarquía de como ve los datos.

Drill Down (Taladrar). Es poder tomar cualquier dimensión de datos y expandir o desglosarlo en componentes más pequeños.

RDBMS. Sistema de gestión de bases de datos relacionales. Es un programa que sirve para crear, diseñar y manipular bases de datos.

Información multidimensional. La información se dice que es multidimensional cuando está organizada de forma que se pueda refernciar desde multiples variables. Así como las tablas bidimensionales, la información es referenciada por dos variables (filas y columnas) en tablas de más dimensiones se refleja información de la relación entre más variables..

ROLAP . La etiqueta ROLAP se aplica a los productos OLAP que toman una base de datos relacional como soporte para la extracción de la información.

MOLAP. La etiqueta MOLAP se aplica a los productos OLAP que toman una base de datos multidimensional como soporte para la extracción de la información.

OLTP-to-OLAP. Las siglas OLTP identifican On Line Transactional Procesing, que son los procesos que tradicionalmente se realizan en la empresa, y de los que se va almacenando la información. Etiqueta el proceso de migración de datos desde el primer tipo de procesos al segundo, esta migración es un elemento crucial dentro de las aplicaciones OLAP, ya que la automatización del proceso y la celeridad del mismo garantiza un sistema actualizado y sencillo.

Hipercubo de datos. Un hipercubo dentro del ámbito de sistemas de información y bases de datos, significa un objeto multidimensional, en el que cada dimensión representa una variable, de forma que en este objeto se obtienen los valores relacionados de todas variables, asociadas a las dimensiones, simultáneamente.

Base de datos multidimensional. Estos tipos de bases de datos usan estructuras para almacenar los datos que tienen la capacidad de relacionar más de dos variables, como ocurre con las tablas en las bases de datos tradicionales. En las bases de datos multidimensionales se usan cubos de datos, hipercubos de datos o multicubos de datos, todas estas estructuras tienen la propiedad que caracteriza a estas bases de datos.

E.I.S. Sistema de información para ejecutivos. Sistema basado en computador concebido para apoyar el trabajo de los directivos de mayor nivel de una empresa.

Bases de datos relacionales. Sistema de almacenamiento de datos basado en un conjunto de tablas unidas mediante relaciones de diversos tipos (binarias, ternarias, agregaciones, generalizaciones etc.).

Multicubo. Es una estructura de organización de datos que distribuye los mismos en numerosos cubos de varias dimensiones. Es decir, en subestructuras en las que se pueden relacionar los datos mediante varias dimensiones.

HOLAP . Es una herramienta OLAP que convina las tecnologías de ROLAP y MOLAP. Es decir, el soporte de almacenamiento de los datos y el motor de generación de las distintas vistas es una combinación de los de ambastecnologías.

Data Warehouse. Es un almacen de datos históricos, en el que se basa una herramienta OLAP para procesar información y elaborar informes y vistas que relacionen los datos, para de este modo ayudar a la toma de decisiones.

SQL. Select Query Language. Es un lenguaje orientado a la creación de consultas de bases de datos.

Arquitectura cliente/servidor. Es un modelo de trabajo cooperativo que se usa en los entornos de red, de forma que una aplicación principal (servidor) se encarga de proporcionar determinados servicios a otros oredenadores o aplicaciones (clientes).

BIBLIOGRAFÍA

- Borrajo Díaz Lourdes, Data Mining, **TI MAGAZINE, Tecnologías de Información.** http://www.timagazine.net , Diciembre 1999, fecha de consulta: 23-05-00, 5:38 p.m.

- Latin Retail Systems, DSS, http://www.latinretail.com/DSS.html, Junio 2000, fecha de consulta: 10-07-00, 6:40 p.m.

- Presser Carner, Cynthia, Data Mining, http://www.monografias.com/trabajos/Data_Mining/Data_Mining.shtml, fecha de consulta: 16-06-00 3:20 p.m.

- Presser Carner Cynthia, Data Mining, http://www.monografias.com/trabajos3/ctrolgestion/ctrolgestion.shtml, fecha de consulta: 16-06-00 4:30 p.m.

- Adrians Peter & Dolf, DATA MINING, De. Addison Wesley, 1996.

- Harjinder S. Gill & Prakash C. Rao, DATA WAREHOUSING, La Integración de Información para la mejor toma de decisiones, Ed. Prentice Hall, 1a. edición, 1996.

- Universidad de Alicante, Depto. De Economía Financiera, Contabilidad y Marketing, Qué son las herramientas OLAP, www.sie.ua.es/webolap/pag2.htm, 12-Ago-1999, fecha de consulta: 15-Junio-2000, 12:44 hrs.

- TREC internet, Redes Neurales Artificiales, http://electronica.com.mx/neural e-mail: neural@electronica.com.mx última actualización: 26-May-1999 12:12:58 PM, fecha de consulta : 14-Jul-2000, 20:30 hrs.

- Larrañaga P., Herriko Unibertsitatea Euskal, Departamento de Ciencias de la Computación e Inteligencia Artificial Universidad del País Vasco http://www.geocities.com/CapeCanaveral/9802/3d5ca100.htm, fecha de consulta: 31-Jul-2000.

- Gómez Flechoso Antonio José, Tesis Doctoral: Inducción de Conocimiento con incertidumbre en Bases de Datos Relacionales Borrosas, Escuela Técnica

Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, http://www.gsi.dit.upm.es/~anto/tesis/html/stateart.html, Madrid, 1998

- EARTHWEB, http://datamation.earthweb.com/dataw/9910mine2.html, fecha de consulta: 10 de agosto de 2000.