



CENTRO DE ESTUDIOS FINANCIEROS

VIRIATO, 52	28010 MADRID	914 44 49 20
PONZANO, 15	28010 MADRID	914 44 49 20
G. DE GRÀCIA, 171	08012 BARCELONA	934 15 09 88
ALBORAYA, 23	46010 VALENCIA	963 61 41 99

www.cef.es

info@cef.es

Índice Tema 7

1. Introducción.
2. Minería de datos. Aplicación a la resolución de problemas de gestión.
 - 2.1. Definición.
3. Fases del proceso de Data Mining.
 - 3.1. Selección y pre-procesamiento de los datos.
 - 3.2. Búsqueda de patrones.
 - 3.3. Interpretación y evaluación.
4. Modelos.
 - 4.1. Modelo de verificación.
 - 4.2. Modelo de descubrimiento.
 - 4.3. Modelo predictivo.
5. Tecnologías utilizadas en la minería de datos.
 - 5.1. Aprendizaje inductivo.
 - 5.2. Estadística.
 - 5.3. Máquina de aprendizaje.
6. Técnicas de la minería de datos.
 - 6.1. De consulta o informe.
 - 6.2. De inteligencia artificial.
 - 6.3. De análisis multidimensional.
7. Algoritmos empleados en Data Mining.
 - 7.1. Algoritmos de clasificación o agrupación.
 - 7.1.1. Métodos de clasificación de datos.
 - 7.1.2. Métodos de abstracción de datos.

- 7.1.3. Aprendizaje de reglas de clasificación.
 - 7.1.4. Algoritmos paralelos.
 - 7.2. Algoritmos de asociación de reglas.
 - 7.3. Algoritmos de análisis secuencial.
- 8. Aplicaciones de la minería de datos.
- 9. Procesamiento analítico en línea (OLAP).
 - 9.1. Definición.
 - 9.2. Bases de datos multidimensionales.
 - 9.2.1. Dimensiones.
 - 9.2.2. Jerarquías.
 - 9.2.3. Series temporales.
 - 9.2.4. Conclusión.
 - 9.3. Características de los sistemas OLAP.
 - 9.4. Tipos.
 - 9.5. Esquemas en estrella e índices bitmap.
 - 9.5.1. Esquema en estrella.
 - 9.5.2. Índices bitmap.
- 10. Elección de una herramienta OLAP.



CENTRO DE ESTUDIOS FINANCIEROS

VIRIATO, 52	28010 MADRID	914 44 49 20
PONZANO, 15	28010 MADRID	914 44 49 20
G. DE GRÀCIA, 171	08012 BARCELONA	934 15 09 88
ALBORAYA, 23	46010 VALENCIA	963 61 41 99

www.cef.es

info@cef.es

TEMA 7

Minería de datos. Aplicación a la resolución de los problemas de gestión. Tecnología y algoritmos. Procesamiento analítico en línea (OLAP).

1. INTRODUCCIÓN.

Tradicionalmente se ha dicho que la información es poder y actualmente esto es más cierto que nunca. La recolección y almacenamiento de datos relativos a productos, clientes, operaciones, transacciones on-line, etc., se ha convertido en un proceso rutinario que está aumentando de forma exponencial la cantidad de datos informatizados a los que se tiene acceso, y sobre los cuales debe basar su estrategia de actuación cualquier organización actual.

Esta sobreabundancia de datos provoca un problema a la hora de tomar decisiones ya que, como la experiencia nos demuestra, es tan difícil elegir la opción acertada cuando no se dispone de suficientes datos sobre un problema como cuando se tienen demasiados: obtener una información completa a partir de datos insuficientes es, obviamente, imposible, pero, igualmente, si el volumen de datos a procesar es tal que resulta imposible tratarlo con las herramientas disponibles en un tiempo razonable, entonces la información obtenida será incompleta o engañosa por no haber tomado en consideración todos los elementos.

En ambos casos la palabra clave es información: no son los datos los que nos van a permitir tomar la decisión acertada, sino la información útil, el significado que podamos extraer de dichos datos. Para lograr extraer este conocimiento útil, es fundamental que se cumplan dos condiciones básicas:

1. Hay que tener suficientes datos sobre los que trabajar y, además, estos datos deben estar organizados en un entorno específicamente diseñado para facilitar su tratamiento por medio de herramientas automatizadas de extracción de información.
2. Debe contarse con herramientas software capaces de tratar grandes cantidades de datos en un tiempo razonable. Estas herramientas se utilizarán para extraer información de las bases de datos, ya sea mediante respuestas a preguntas concretas o mediante la búsqueda de tendencias y relaciones entre los datos disponibles.

Las técnicas y herramientas que se han desarrollado para responder a esta necesidad acuciante de disponer de una información exacta y en un tiempo razonable, para cumplir los objetivos corporativos, se ha convertido en un tema de importancia para cualquier organización. Como solución al problema de tener un entorno de análisis adecuado es lo que se ha dado en llamar almacén de datos (en inglés, Data Warehouse). Como respuesta al problema de extracción de información a partir del almacén de datos hablaremos de las herramientas de minería de datos (Data Mining) y OLAP (On-Line Analytical Processing).

2. MINERÍA DE DATOS. APLICACIÓN A LA RESOLUCIÓN DE PROBLEMAS DE GESTIÓN.

El almacén guarda todos los datos relevantes para una organización y cómo está estructurado para que se pueda extraer información a partir de dichos datos. Sin embargo, sin las herramientas de consulta adecuadas, de nada sirve el almacén. Por muy bien construido que esté, aunque el modelo de datos refleje fielmente la realidad de la organización, el proceso de carga sea perfecto y la base de datos se encuentre soportada por un SGBD y una arquitectura distribuida que ofrezca unos tiempos de respuesta espectaculares, si no se dispone de unas herramientas que extraigan la información que se halla oculta entre la montaña de datos es como si no se tuviera nada.

La minería de datos permite sacar el máximo provecho del almacén de datos, ofreciendo una serie de técnicas y herramientas que automatizan el proceso de extracción de información y significado a partir de los datos que éste contiene.

Dado que el objetivo último de la gestión de datos corporativos es ofrecer información de calidad a la dirección, cuanto más eficiente sea este proceso de minería, mayor será en cantidad y en calidad la información disponible para soportar la toma de decisiones.

2.1. DEFINICIÓN.

Aunque no hay una única definición de minería de datos, se puede decir que es un conjunto de técnicas y procesos de análisis de datos que permite extraer información de bases de datos mediante la búsqueda automatizada de patrones y relaciones.

Como definición más formal se puede tomar la siguiente:

La minería de datos consiste en la búsqueda de relaciones y patrones globales que se hallan presentes en las grandes bases de datos pero que están «ocultos» entre el gran volumen de datos existente. Estas relaciones representan un conocimiento útil sobre los objetos de la base de datos y la realidad que representan [Marcel Holshemier & Arno Siebes (1994)].

Pese a la variedad de definiciones que se pueden encontrar, todas ellas tienen unos puntos en común:

- Es necesario disponer de unas bases de datos o, mejor aún, de un almacén de datos, sobre los cuales realizar el proceso de minería.
- El proceso de minería debe ser automático, en la mayor medida posible, debido a los grandes volúmenes de datos que se deben analizar.
- Los resultados obtenidos deben representar conocimiento útil y no evidente a primera vista.

Teniendo presentes estas características, el proceso de minería de datos (Data Mining) se divide en:

- Fases: diferentes etapas del proceso.
- Modelos: modelos a los que se ajustan las herramientas de Data Mining.
- Tecnologías: cuáles son las tecnologías base del Data Mining.
- Algoritmos: algoritmos empleados en el Data Mining.

3. FASES DEL PROCESO DE DATA MINING.

El proceso de minería de datos se puede dividir en las siguientes fases:

3.1. SELECCIÓN Y PRE-PROCESAMIENTO DE LOS DATOS.

Si se va a trabajar sobre un almacén de datos (Data Warehouse), este paso no es necesario puesto que los datos ya están preparados para ser utilizados, en otro caso hay que realizar prácticamente las mismas tareas que conducen al establecimiento de un almacén.

Concretamente, en esta fase se decide sobre qué datos se va a trabajar, tanto desde el punto de vista físico (ficheros planos, bases de datos relacionales, servicios on-line, etc.) como desde el punto de vista lógico (datos personales, de ventas, de productos, etc.)

También se deben depurar los datos en bruto antes de poder comenzar el proceso de extracción de información. Como ejemplo típico de depuración está la eliminación de datos irrelevantes, la unificación de criterios de representación que pueden no ser los mismos en todas las fuentes de datos a las que se va a acceder o la eliminación de redundancias y duplicados.

3.2. BÚSQUEDA DE PATRONES.

Ésta es la fase en la que la herramienta de minería de datos, ya sea con la asistencia de un operador humano o de forma autónoma, analiza los datos ya preparados para extraer significado e información. El resultado de la búsqueda será un informe que debe ser analizado por un operador humano en la siguiente fase.

3.3. INTERPRETACIÓN Y EVALUACIÓN.

Esta fase final consiste en la interpretación de los resultados producidos en la fase de búsqueda. Como resultado de esta interpretación se puede llevar a cabo alguna de las siguientes acciones:

- Volver de nuevo a la búsqueda de patrones, con consultas más refinadas si se está usando el modelo de verificación.
- Utilizar los resultados como entrada para un sistema de soporte a la decisión.
- Emplearlos directamente en la preparación de un informe para la dirección.

4. MODELOS.

4.1. MODELO DE VERIFICACIÓN.

Éste es el modelo más parecido al proceso tradicional de extracción de información basado en lenguajes de consulta a bases de datos (por ejemplo, SQL). Su principal característica es que no extrae información nueva, sino que, basándose en los datos del almacén, verifica la validez de las afirmaciones que se le presentan.

El proceso comienza por el establecimiento de una hipótesis por parte del usuario. Éste, a continuación, solicita a la herramienta que verifique su validez. Una vez recibida la respuesta, el usuario puede refinar o detallar la hipótesis, preparar unas preguntas más específicas y solicitar una nueva verificación. De esta manera se consigue un proceso iterativo dirigido por el operador humano.

La utilización de herramientas de este tipo presenta la desventaja de que si al usuario no se le ocurre realizar una pregunta clave, o no ve una relación importante entre diferentes elementos de la base de datos, la herramienta por sí sola carece de iniciativa para investigar por su propia cuenta.

4.2. MODELO DE DESCUBRIMIENTO.

En este segundo modelo se utiliza la herramienta de minería para descubrir nueva información que no estaba anteriormente en el almacén de forma explícita. Según este modelo es la propia herramienta la que se plantea sus propias preguntas, sin necesidad de que el usuario establezca hipótesis o realice preguntas concretas, aunque éste puede intervenir para guiar los caminos a explorar.

Habitualmente esta búsqueda se dirige hacia la categorización de los registros en grupos para detectar patrones aplicables o extraer relaciones implícitas en los datos. Este tipo de búsqueda es útil, por ejemplo, a la hora de clasificar clientes en función de sus hábitos de consumo, estatus social, etc.

También es común la búsqueda de elementos extraños o fuera de la norma. Esto es de utilidad para detectar posibles fraudes como, por ejemplo, la existencia de personas con una baja renta declarada que, por otra parte, dispongan de un coche deportivo de importación.

4.3. MODELO PREDICTIVO.

El objetivo de este modelo de minería consiste en la realización por parte de la herramienta de predicciones sobre el comportamiento futuro de variables a partir de los patrones existentes en los datos.

En este caso el usuario indica sobre qué variables se quiere obtener la predicción y el sistema proporciona la respuesta. Esta respuesta la puede proporcionar explicando cómo la consiguió, lo cual a su vez puede ser una información tan valiosa como la respuesta en sí misma, o sin explicarlo.

En el caso de que no haya suficientes datos para realizar una predicción fiable se puede optar por dos caminos alternativos: no obtener predicción alguna (modelo predictivo restringido) u obligar a la realización de una predicción de menor fiabilidad (modelo predictivo no restringido).

5. TECNOLOGÍAS UTILIZADAS EN LA MINERÍA DE DATOS.

Las herramientas de Data Mining, en cuanto a extracción de datos y búsqueda de patrones, se basan en tres tipos de tecnologías: el aprendizaje inductivo, la máquina de aprendizaje y el estudio de estadísticas.

5.1. APRENDIZAJE INDUCTIVO.

Inducción es la inferencia de información a partir de datos. El aprendizaje inductivo es el modelo de construcción de procesos donde las bases de datos son analizadas para buscar patrones. Objetos similares son agrupados en clases y las reglas son formuladas de tal forma que hacen posible predecir la clase de los objetos invisibles o inadvertidos. Este proceso de clasificación identifica clases tales que cada una de ellas tiene un único modelo de valores que forma su descripción de clase. La naturaleza del entorno es dinámica, de aquí que el modelo deba ser adaptativo, debe ser capaz de aprender.

Generalmente esto es posible para un número pequeño de propiedades que caracterizan el objeto, así que se puede hacer abstracción en aquellos objetos que satisfacen un subconjunto de propiedades que son proyectadas con la misma representación interna.

El aprendizaje inductivo, donde el sistema infiere conocimientos a partir de la observación de su entorno, tiene dos principales estrategias:

- Aprendizaje supervisado: aprendizaje a partir de ejemplos, donde el maestro ayuda al sistema a construir un modelo mediante la definición de clases y de ejemplos de cada clase. El sistema tiene que definir una descripción de cada clase: propiedades comunes de los ejemplos. Una vez que la descripción ha sido formulada, descripción y clase forman una regla de clasificación que puede ser utilizada para predecir la clase de objetos previamente invisibles.
- Aprendizaje no supervisado: aprendizaje mediante observación y descubrimiento. El sistema de Data Mining es provisto de objetos, pero no de clases que estén definidas. De esta forma, tiene que observar los objetos y reconocer los patrones (descripción de la clase) por sí mismo. El resultado de este sistema es un conjunto de descripciones de clases, una por cada una de las clases descubiertas en el entorno.

Inducción es por tanto extraer patrones o modelos de referencia. La calidad del modelo producido por el aprendizaje inductivo es tal que el modelo puede ser usado para predecir el resultado de situaciones futuras, no sólo para estados encontrados sino para estados invisibles que puedan ocurrir. El problema es que la mayor parte de los entornos tienen diferentes estados, cambios internos, y no siempre es posible verificar un modelo probándolo en todas las posibles situaciones.

Dado un conjunto de ejemplos, el sistema puede construir múltiples modelos, algunos de los cuales pueden ser más simples que otros. Los modelos más simples serán probablemente correctos, ya que si hubiera múltiples explicaciones a un fenómeno en particular, tiene sentido elegir las más simples porque es la que con más probabilidad captura la naturaleza del fenómeno (Ockhams).

5.2. ESTADÍSTICA.

La estadística tiene un sólido fundamento teórico pero los resultados pueden ser abrumadores y difíciles de interpretar, requiere la guía del usuario y saber dónde y cómo se analizan los datos. Los sistemas de Data Mining permiten que el conocimiento de los expertos de los datos y las técnicas avanzadas de análisis de los ordenadores trabajen conjuntamente.

Sistemas de análisis estadísticos como SAS y SPSS han sido utilizados por los analistas para detectar patrones inusuales y explicarlos usando técnicas estadísticas de modelos lineales.

5.3. MÁQUINA DE APRENDIZAJE.

La máquina de aprendizaje es la automatización de un proceso de aprendizaje y, el aprendizaje es el equivalente a la construcción de reglas basadas en la observación de los estados y transiciones del entorno. Éste es un campo muy amplio que incluye no sólo el aprendizaje a partir de ejemplo, sino que también refuerza el mismo.

Un algoritmo de aprendizaje tiene como entrada el conjunto de datos y su información adicional, y como salida un estamento: un concepto representando los resultados del aprendizaje de la entrada. La máquina de aprendizaje examina los ejemplos previos y sus resultados, y aprende cómo reproducir éstos y a hacer generalizaciones sobre nuevos casos.

Generalmente un sistema de máquina de aprendizaje no sólo utiliza observaciones sencillas de su entorno sino un conjunto finito llamado conjunto de ensayo. Este conjunto contiene ejemplos en un formato que la máquina pueda leer.

6. TÉCNICAS DE LA MINERÍA DE DATOS.

Algunas de las técnicas más comúnmente utilizadas para la extracción de información son las siguientes:

- De consulta o informe.
- De inteligencia artificial.
- De análisis multidimensional.

6.1. DE CONSULTA O INFORME.

Ésta es la forma tradicional de obtener información a partir de bases de datos relacionales. Consiste en la utilización de herramientas que facilitan la construcción de consultas SQL mediante interfaces gráficos, lanzan dicha consulta y a continuación presentan los resultados en forma de tablas, diagramas o gráficos. Adicionalmente pueden utilizar técnicas matemáticas y estadísticas para analizar los datos obtenidos.

Estas técnicas son muy apropiadas si se va a utilizar el modelo de minería de verificación. Su principal ventaja es que son de eficacia probada, trabajan sobre las bases de datos relacionales ya existentes y además es muy sencillo encontrar herramientas amigables al usuario que las soporten.

6.2. DE INTELIGENCIA ARTIFICIAL.

Tomadas de campos de la inteligencia artificial tales como los sistemas expertos, el aprendizaje automático, la visión por ordenador o la teoría de juegos. Utilizan estructuras de datos y algoritmos basados en árboles de decisión, redes neuronales, técnicas de agrupamiento o clustering y lógica difusa.

Estas técnicas son especialmente adecuadas para herramientas de minería que utilizan los modelos predictivo y de descubrimiento, ya que son muy buenas en la detección de patrones. Es fácil encontrarlas formando parte de lo que se ha dado en llamar agentes inteligentes.

6.3. DE ANÁLISIS MULTIDIMENSIONAL.

Basadas en la utilización de bases de datos multidimensionales. Estas bases almacenan los datos de forma parecida a como lo hace una hoja de cálculo aunque, a diferencia de éstas, es común que utilicen más de dos dimensiones.

Las técnicas multidimensionales, o técnicas de procesamiento analítico en línea (OLAP) son muy buenas para cruzar los datos de múltiples formas y con distintos niveles de agregación. Por ejemplo, dado el cubo n-dimensional que forma la base de datos, obtener vistas bidimensionales de los datos (por ejemplo ventas-región, producto-ventas, etc.) de forma dinámica es un proceso mucho más rápido y sencillo de lo que lo sería utilizando un SGBDR.

7. ALGORITMOS EMPLEADOS EN DATA MINING.

Los algoritmos de Data Mining se clasifican en tres grandes grupos:

- Agrupación o clasificación.
- Reglas de asociación.
- Secuencia de análisis.

7.1. ALGORITMOS DE CLASIFICACIÓN O AGRUPACIÓN.

Con estos algoritmos se analiza un conjunto de datos y se genera un conjunto de reglas agrupadas que puede ser utilizadas para clasificar futuros datos. Por ejemplo, Uno puede clasificar enfermedades y proporcionar los síntomas que describen cada clase y subclase.

Uno de los problemas más importantes del Data Mining es el aprendizaje de las reglas de clasificación, que involucra encontrar reglas que dividan los datos en clases predefinidas. En el dominio del Data Mining donde se trabaja con millones de registros y gran cantidad de atributos, el tiempo de ejecución de los algoritmos existentes puede ser prohibitivo, especialmente para aplicaciones interactivas.

7.1.1. Métodos de clasificación de datos.

Estos algoritmos son lo que más éxito han tenido en sus diferentes campo de aplicación, especialmente en los basados en la máquina de aprendizaje. Estos algoritmos se pueden dividir:

- Algoritmos estadísticos: los sistemas de análisis estadísticos han sido utilizados por los analistas para detectar patrones inusuales. También pueden explicarlos utilizando modelos estadísticos basados en técnicas lineales.

- Redes neuronales: las redes artificiales neuronales imitan la capacidad de encontrar patrones del cerebro humano, de aquí que algunos investigadores hayan sugerido aplicar algoritmos de redes neuronales para la asignación de patrones.
- Algoritmos genéticos: son técnicas de optimización que utilizan procesos como combinaciones genéticas, mutaciones y selección natural. El diseño está basado en conceptos de evolución natural.
- Método del vecino más próximo: es una técnica que clasifica cada registro de un conjunto de datos basándose en combinaciones de las clases de los k registros más similares a él en un histórico.
- Regla de inducción: extracción de reglas «si-entonces» a partir de datos de relevancia estadística.
- Visualización de datos: interpretación visual de relaciones complejas en sistemas multidimensionales (OLAP).

7.1.2. Métodos de abstracción de datos.

Muchos de los algoritmos existentes sugieren datos extra de prueba antes de clasificarlos en varias clases. Hay varias opciones para hacer la abstracción antes de la clasificación: un conjunto de datos puede ser generalizado a un nivel de abstracción mínimo, nivel de abstracción intermedio, nivel de abstracción alto.

Un nivel de abstracción demasiado bajo puede originar demasiadas clases dispersas o árboles de clasificación con demasiadas ramas, dificultando una interpretación semántica concisa.

Un nivel de abstracción demasiado alto puede provocar una pérdida de propiedad de clasificación.

7.1.3. Aprendizaje de reglas de clasificación.

Estos algoritmos comprenden encontrar reglas o árboles de decisión que dividan los datos en clases predefinidas. Para cualquier dominio de un problema real el conjunto de posibles árboles de decisión es demasiado amplio para ser investigado exhaustivamente. De hecho, la complejidad de cálculo de encontrar el árbol de decisión óptimo es muy complicada. Este tipo de algoritmos disminuyen esa complejidad.

Basados en algoritmos de inducción utiliza el método de Hunt como base. Construye un árbol de decisión a partir de un conjunto de casos de entrenamiento.

Algoritmos que se clasifican dentro de este grupo son:

- ID3: es un algoritmo de construcción de árbol de decisión que determina la clasificación de los objetos sometiendo a prueba los valores de sus propiedades. Construye el árbol de arriba abajo, comenzando a partir de un conjunto de objetos y una especificación de sus propiedades. Para cada nodo del árbol, una propiedad es analizada y el resultado se usa para dividir el conjunto de objetos. Este proceso se ejecuta de forma recursiva hasta que el conjunto es un subárbol homogéneo respecto a los criterios de clasificación; contiene objetos pertenecientes a la misma categoría. Este grupo de objetos se convierte en una hoja del nodo. Para cada nodo, la propiedad so-

metida a ensayo se elige basándose en criterios teóricos de información para buscar maximizar la ganancia de información y minimizar la entropía. En términos más simples, la propiedad analizada se utiliza para dividir el conjunto candidato en subconjuntos más homogéneos.

- C4.5: este algoritmo genera árboles de decisión-clasificación para un conjunto de datos mediante división recursiva de los datos. El algoritmo considera todos los ensayos en los que se pueden dividir los datos y selecciona un ensayo que proporcione la mejor ganancia de información.
- SLIQ: es un árbol de decisión clasificador diseñado para clasificar datos de ensayo de gran tamaño. Utiliza una técnica de pre-ordenación en la fase de crecimiento del árbol.

7.1.4. Algoritmos paralelos.

La mayoría de los algoritmos existentes utilizan técnicas heurísticas para gestionar la complejidad de cálculo, ya que ésta es muy elevada especialmente cuando el número de atributos y el número de datos de ensayo es alto.

Los algoritmos paralelos han sido sugeridos por muchos grupos como base para la minería de datos. La idea básica es: se eligen inicialmente N datos de ensayo distribuidos aleatoriamente por P procesadores de tal forma que $N = P$. Todos los procesadores cooperan para expandir el nodo raíz del árbol de decisión, lo que puede ser hecho en tres pasos:

- Cada procesador recolecta información de distribución de la clase de los datos locales.
- Cada procesador intercambia esta información.
- Cada procesador puede simultáneamente calcular la ganancia de entropía de los atributos y encontrar el mejor atributo para dividir el nodo raíz.

Hay dos aproximaciones para estos algoritmos:

- Aproximación de construcción síncrona del árbol (Synchronous Tree Construction Approach): donde todo el conjunto de procesadores expanden síncronamente un nodo del árbol de decisión al mismo tiempo.
- Aproximación de construcción del árbol por división (Partitioned Tree Construction Approach): donde cada nodo nuevo generado es expandido por un subconjunto de procesadores que ayudan a la expansión del nodo padres.

7.2. ALGORITMOS DE ASOCIACIÓN DE REGLAS.

Una regla de asociación es aquella que implica relaciones de asociaciones válidas entre un conjunto de objetos de una base de datos, tales como: «ocurren juntos» o «la ocurrencia de uno implica la de otro». En este proceso se descubre un conjunto de reglas de asociación en múltiples niveles de abstracción a partir de los conjuntos de datos relevantes de la base de datos. Por ejemplo, se puede descubrir un conjunto de síntomas que normalmente ocurren conjuntos en ciertos tipos de enfermedades y estudiar posteriormente las razones que hay detrás.

Dado un conjunto de transacciones, donde cada transacción es un conjunto de literales llamados items una regla de asociación es una expresión de la forma $X \rightarrow Y$ donde X e Y son conjuntos de items. El significado intuitivo de estas reglas es aquella transacción de la base de datos que contenga X tiende a contener Y . Un ejemplo de una regla de asociación sería: «30 por 100 de las transacciones que contienen "cerveza" también contienen "pañales"; 2 por 100 de todas las transacciones contienen ambos de estos items». En este caso se llama al 30 por 100 confianza de la regla, y al 2 por 100 se le llama el soporte de la regla.

El problema es encontrar todas las reglas de asociación que satisfagan el mínimo soporte y la confianza mínima especificadas por el usuario, que actúan como restricciones del sistema.

Algunos de los algoritmos utilizados son:

- Algoritmo *a priori*: desarrollado para grandes bases de datos transaccionales.
- Algoritmos distribuidos/paralelos: distribuyen un gran sistema de bases de datos en sistemas distribuidos para facilitar el uso del algoritmo.

7.3. ALGORITMOS DE ANÁLISIS SECUENCIAL.

Con estos algoritmos, se descubren modelos que se suceden en una secuencia. Esto tiene que ver con los datos que aparecen en transacciones separadas, a diferencia de los datos que aparecen en la misma transacción en el caso de asociación. Por ejemplo: si un comprador compra un elemento A la primera semana del mes, posteriormente compra el elemento B en la segunda semana del mes, etc.

Cada dato de entrada es un conjunto de secuencias llamada secuencias de datos. Cada secuencia de datos es una lista ordenada de transacciones, donde cada transacción es un conjunto de literales o términos (items).

Hay una transacción temporal asociada con cada transacción. Un modelo de secuencia también consiste en una lista de conjuntos de términos. El problema es encontrar todos los patrones de secuencia con un soporte mínimo especificado por el usuario, donde el soporte de un patrón secuencial es el porcentaje de la secuencia de datos que contiene ese patrón.

Un ejemplo de un modelo de este tipo es que los clientes usualmente alquilan: «La Guerra de las Galaxias», posteriormente «El Imperio Contraataca» y luego «El Retorno del Jedi». Hay que notar que no todos estos alquileres necesitan ser consecutivos. Habrá clientes que alquilarán otros vídeos entre los mencionados, pero que también apoyan esa secuencia. También hay casos donde los elementos de una secuencia no tienen por qué ser elementos sencillos.

8. APLICACIONES DE LA MINERÍA DE DATOS.

La minería de datos tiene una gran variedad de campos de aplicación, algunos de los cuales los comentamos a continuación:

- Estudios de ventas y mercado: consiste en identificar patrones de compra de los consumidores, encontrar asociaciones entre sus distintas características demográficas y poder predecir su respuesta a las distintas campañas de correo comercial.

- Banca: en este caso la minería de datos se utiliza para detectar los patrones de personas que usan de forma fraudulenta las tarjetas de crédito así como para identificar la lealtad de los clientes. Poder predecir los clientes que con probabilidad cambiarán su tarjeta de crédito. Determinar el gasto de los clientes que utilizan la tarjeta de crédito y agruparlos por determinadas características. Encontrar correlaciones ocultas entre diferentes indicadores financieros.
- Sanidad y Seguridad Social: identificar patrones de comportamiento de grupos de riesgo. Identificar comportamientos fraudulentos (para pago de cotizaciones y cobro de subvenciones y pensiones).
- Transporte: determinar distribución de horarios entre las distintas salidas de material. Analizar los patrones de carga.
- Medicina: caracterizar el comportamiento del paciente para predecir cuándo necesitará nuevas consultas. Identificar terapias médicas de éxito para distintas enfermedades.
- Sistemas policiales: relacionar hechos delictivos con personas y lugares y ser capaz de predecir el comportamiento criminal y sus objetivos.

9. PROCESAMIENTO ANALÍTICO EN LÍNEA (OLAP).

9.1. DEFINICIÓN.

El acrónimo OLAP significa Procesamiento Analítico En-línea (On-Line Analytical Processing) y se utiliza para hacer referencia a sistemas y herramientas de minería de datos que utilizan técnicas multidimensionales para la extracción y el análisis de los datos.

Según E.F. Codd, que fue quien acuñó el término, OLAP es «la síntesis, el análisis y la consolidación dinámica de grandes volúmenes de datos multidimensionales».

Siempre que se habla de tecnología o herramientas OLAP el adjetivo más utilizado es «multidimensional», ya sea para referirse a los datos, a su estructura, a la base de datos que se emplea o a casi cualquier otro aspecto del OLAP. Esta caracterización llega hasta el punto de identificar el OLAP y las bases de datos multidimensionales como una misma cosa. Aunque, indudablemente, ambas tecnologías están relacionadas, la utilización de OLAP no implica necesariamente la utilización de bases de datos multidimensionales.

Para poder distinguir la diferencia entre uno y otro concepto, vamos a hacer un paréntesis en la explicación de OLAP para ver con algo de detalle qué es exactamente una base de datos multidimensional, en qué se diferencia de una base de datos relacional y qué ventajas reporta su utilización.

9.2. BASES DE DATOS MULTIDIMENSIONALES.

La idea básica empleada por las bases de datos multidimensionales (BDM en adelante) es muy sencilla: en vez de utilizar tablas para almacenar los datos, como se hace en una base de datos relacional (BDR en adelante), emplea matrices. Es algo parecido a utilizar una hoja de cálculo para el tratamiento de datos, sólo que, habitualmente, se utilizarán más de dos dimensiones y se dispondrá de otras capacidades adicionales. Vamos a ver a continuación un ejemplo sencillo de BDM para aclarar el concepto.

En las bases de datos multidimensionales los conceptos básicos son:

- Dimensiones.
- Jerarquías.
- Series temporales.

En los siguientes apartados se verán con unos ejemplos cada uno de estos conceptos:

9.2.1. Dimensiones.

EJEMPLO 1:

Supongamos que queremos implementar una sencilla base de datos para almacenar la cantidad de dinero que se gasta en el pago de las pensiones atendiendo al tipo de pensión y a la comunidad autónoma en que se paga.

En el caso de que hubiera dos tipos de pensiones, se podría establecer una BDM con una estructura similar a la de una hoja de cálculo, empleando tantas filas como tipos de pensiones y tantas columnas como comunidades autónomas. El gasto correspondiente a cada comunidad y pensión se almacenaría en la celdilla correspondiente, tal como se muestra en la siguiente figura:

	CA1	CA2	CA3	CA4	CA5	CA6	CA7	CA8	CA9	CA10	CA11	CA12	CA13	CA14	CA15	CA16	CA17
P1																	
P2																	

El equivalente relacional sería una tabla de tres columnas y 34 filas (suponiendo que en todas las comunidades se pagan ambos tipos de pensiones) como la que se muestra a continuación:

TIPO PENSIÓN	COMUNIDAD AUTÓNOMA	GASTO
P1	CA1	G1
P1	CA2	G2
...

En este ejemplo sencillo, el espacio de almacenamiento utilizado en ambos casos es el mismo, pero, ¿qué ocurre con los tiempos de acceso a la información?

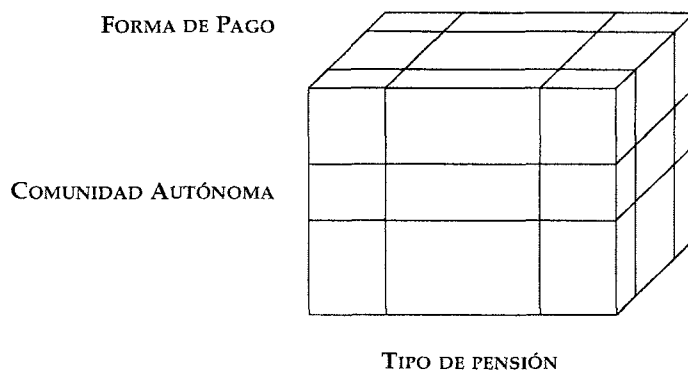
Si se quiere acceder al gasto en un tipo de pensión y una comunidad autónoma determinados (recuperar una única tupla), el tiempo de acceso será similar siempre que la tabla relacional esté ordenada o tenga definido un índice.

En cambio, si lo que se quiere obtener es el gasto en pensiones del tipo 1 para todas las comunidades, entonces el tiempo de respuesta de la BDM será mucho más rápido, ya que sólo tiene que sumar una fila de la matriz (17 sumas). En cambio, la BDR debe recorrer todos los registros de la tabla para localizar aquellos que cumplan la condición definida (pensión tipo 1), lo cual implica el procesamiento de los 34 registros.

Por supuesto, esta diferencia se puede acortar mediante el empleo de índices en la BDR tanto para pensiones como para comunidades, pero en casos reales con tablas de más columnas, no es factible establecer índices para todos los casos posibles de búsqueda, por lo que la ventaja de la BDM se mantiene.

EJEMPLO 2:

Siguiendo con el ejemplo anterior, supongamos que también es necesario almacenar la forma de pago de las pensiones y que dicha forma de pago puede ser por talón, en efectivo o mediante transferencia. En este caso la BDM tendría el siguiente aspecto:



Aunque para mejorar la claridad del dibujo no aparecen todas las particiones del cubo, es evidente la estructura utilizada: se emplea cada una de las tres dimensiones del cubo para representar cada uno de los campos que se utilizaría en el modelo relacional. Las celdas resultantes se emplean para almacenar el gasto para cada tripleta (CA, TP, FP).

El equivalente relacional consistiría en una tabla de cuatro columnas y 102 registros, en la que de nuevo sería más costoso, computacionalmente hablando, realizar consultas de agregados (totales) que el modelo multidimensional.

Estos dos ejemplos han servido para comprender en qué consiste el almacenamiento en bases de datos multidimensionales, pero eso no es todo, ya que otro aspecto fundamental de las BDM es la posibilidad de jerarquizar las dimensiones. Vamos a ver esto con otro ejemplo.

9.2.2. Jerarquías.

Refinando más el ejemplo de las pensiones, supongamos que, además de conocer el gasto por comunidades, se quiere saber también por localidades dentro de cada comunidad.

La manera inmediata de representar esto consiste en añadir una nueva dimensión para crear un hipercubo de cuatro dimensiones. Sin embargo esta solución no es eficiente, ya que, para cada fila de cada localidad, sólo una de las celdillas contendría valor. Dicha celdilla sería la correspondiente a la comunidad a la que perteneciera la localidad (por ejemplo, para Jerez, sólo la celdilla correspondiente a la Comunidad Autónoma Andaluza podría tener algún dato almacenado).

Con esta estructura se malgasta mucho espacio de almacenamiento en celdillas que jamás van a contener datos, por tanto es necesario utilizar algún mecanismo que lo evite. La solución a este problema consiste en crear una jerarquía de niveles en cada dimensión para representar los diversos grados de detalle. Si se dispone de este mecanismo, la solución al caso de las localidades sería tan simple como jerarquizar la dimensión de las comunidades autónomas, estableciendo las localidades como el escalón inferior de la jerarquía. Para ofrecer esta alternativa, el gestor debe ser capaz, al operar con las celdillas, de reconocer si el valor almacenado corresponde a una comunidad o a una localidad, de forma que al hallar totales o realizar cualquier otra operación, no mezcle valores correspondientes a diferentes niveles jerárquicos.

Por supuesto, el concepto de jerarquía es extensible a más de dos niveles, por lo que se puede afinar el grado de detalle obtenido al realizar las consultas sin más que establecer los niveles de jerarquía adecuados. A este proceso de excavación a través de jerarquías para obtener mayores grados de detalle se a lo que en inglés se llama «drill-down» y es un término muy utilizado al hablar de BDM y OLAP.

9.2.3. Series temporales.

Una de las características de los almacenes de datos es que guardan tanto información actual como histórica. Esta necesidad de guardar datos en función del tiempo, hace que una de las dimensiones más habituales en cualquier BDM sea el propio tiempo.

En nuestro ejemplo, si se quiere guardar un registro de la evolución de los pagos de las pensiones, se podría añadir una nueva dimensión que representara el tiempo. No obstante, como ya hemos visto anteriormente, añadir nuevas dimensiones a la BDM hace que el espacio de almacenamiento necesario crezca muy rápidamente.

Además, en el caso de dimensionar el tiempo se nos presenta otra complicación añadida: ¿qué unidad de medida tomamos para hacer las divisiones? Si elegimos, por ejemplo, meses entonces cualquier dato que introduzcamos deberá hacer referencia obligatoriamente a un mes, ni a un día ni a dos meses ni a un año, sino exactamente a un mes. Esta limitación se podría evitar de dos maneras: jerarquizando la dimensión tiempo o utilizando una unidad de medida más fina.

En el caso de jerarquizar el tiempo, los problemas son tanto el incremento del espacio ocupado como la programación y el tratamiento necesario para introducir los intervalos de fechas en la base de datos.

En el caso de usar una unidad de medida más fina, como por ejemplo un día, surge un problema similar de tratamiento, ya que se debería pasar todos los intervalos a días, lo cual añadiría una carga extra de conversión tanto al introducir datos como al extraerlos.

La solución a este problema consiste en disponer de una BDM que trate las series temporales como una dimensión. Disponiendo de esta capacidad, cada celdilla no almacenaría como valor un único número, sino que guardaría un conjunto de ellos para representar la información histórica. Al ofrecer este nuevo tipo de datos, el gestor de la base de datos debe ser capaz de trabajar con calendarios, convertir entre diferentes intervalos de tiempo, distinguir días festivos y laborables, etc.

9.2.4. Conclusión.

La utilización de BDM ofrece una innegable ventaja sobre las BDR siempre que se vaya a trabajar sobre datos agregados, totales, subtotales, etc. También son superiores a la hora de trabajar con series temporales, obtener vistas de unos datos en función de otros (vistas bidimensionales del hipercubo que forma la BDM) y manejar diversos grados de detalle. En resumen, son unas bases de datos adecuadas para el estudio a alto nivel de los datos, al ofrecer una mayor flexibilidad y rapidez de acceso para el análisis de los mismos.

Por otra parte, si lo que se quiere es acceder a un dato individual básico como puede ser el sueldo de una persona concreta o lo que se ha gastado en pensiones en Jerez el mes de Marzo, la ventaja de las BDM desaparece en favor de las BDR. Éstas son capaces de recuperar un dato individual con la misma eficiencia que las multidimensionales, suelen ser capaces de almacenar mayor cantidad de registros y además, dada su utilización masiva en sistemas OLTP, están optimizadas para la inserción de registros y el control concurrente de usuarios.

Hay que tener claro que la utilización de ambos tipos de bases de datos no es excluyente. De hecho no es infrecuente el caso en que se utiliza una BDR para almacenar los datos del nivel más bajo de la jerarquía de una BDM, de forma que si se desea obtener un dato básico, se excava a través de la jerarquía multidimensional hasta acceder a la BDR. Vamos ahora a estudiar ahora los sistemas OLAP, viendo cuál es su relación con las bases de datos multidimensionales y cómo pueden aprovechar sus capacidades.

9.3. CARACTERISTICAS DE LOS SISTEMAS OLAP.

Las características básicas de los sistemas OLAP son las siguientes:

- Ofrecen una visión multidimensional y jerarquizada de los datos.
- Son capaces de analizar tendencias a lo largo de períodos de tiempo.
- Pueden presentar vistas de un número reducido de dimensiones elegido por el usuario.
- Permiten ahondar en la jerarquía de los datos para acceder a los de más bajo nivel.
- Son interactivos y soportan múltiples usuarios concurrentemente.

Resulta claro, vistas sus características, cómo los sistemas OLAP pueden beneficiarse de las funcionalidades de una BDM de la siguiente forma:

- La visión multidimensional y jerarquizada está explícita en la propia estructura de la base de datos. La herramienta OLAP, que posiblemente esté integrada en la BDM (o viceversa), sólo tiene que ocuparse del manejo del cubo hiperdimensional para extraer los datos conforme a los criterios establecidos por el usuario.
- El estudio de tendencias se puede realizar aprovechando las series temporales de la BDM o, si no se dispone de dicho tipo de datos, realizando las operaciones y conversiones necesarias para manejar el tiempo como una dimensión adicional de la base de datos.

- La presentación de vistas se conoce en la jerga OLAP como «slice and dice» y se podría traducir en algo así como «cortar y trocear». Esta característica de una herramienta OLAP consiste en la capacidad de extraer «rodajas» del hipercubo que forma la BDM. Estas rodajas se extraen tomando un valor fijo para una o varias dimensiones y tomando el hipercubo resultante.

Por ejemplo, en el segundo caso que se vio de BDM, si fijamos el valor de la dimensión comunidad autónoma a Galicia, el resultado es un corte bidimensional del cubo inicial. Este corte tendría el siguiente aspecto:

	EFFECTIVO	TALÓN	TRANSFERENCIA
Forma de pago 1			
Forma de pago 2			

Si la base de datos tuviera más dimensiones, el resultado de proceso tendría tantas dimensiones como las de la base de datos completa menos el número de dimensiones cuyo valor se está fijando. Evidentemente, la presentación de rodajas de más de tres dimensiones no es algo intuitivo, por lo que habitualmente se suelen extraer rodajas de 2 ó 3 dimensiones.

- La capacidad de excavar en los niveles de jerarquía se realiza, de nuevo, aprovechando la propia estructura de la BDM subyacente. En el caso en que se utilice una BDR como escalón inferior de la jerarquía, la herramienta OLAP debe ocuparse de que el acceso a dicho nivel sea transparente para el usuario.
- La interactividad y el soporte de múltiples usuarios simultáneos son capacidades que dependen en gran medida de los tiempos de respuesta del gestor de bases de datos empleado, por lo que se puede utilizar como criterio orientativo a la hora de elegir el producto que se va a adquirir para construir el sistema.

9.4. TIPOS.

Debido a su orientación hacia el manejo de los datos organizados en dimensiones, el entorno natural de trabajo de los sistemas OLAP son las bases de datos multidimensionales. No obstante también pueden trabajar sobre bases de datos relacionales, aunque en este caso sus prestaciones se ven disminuidas. Atendiendo a este criterio, los sistemas OLAP se pueden dividir en dos tipos:

- MOLAP (Multidimensional-OLAP): funcionan sobre bases de datos multidimensionales. Requieren de un esfuerzo inicial previo de modelización y construcción de la base de datos multidimensional y de otro continuo consistente en migrar los datos en formato relacional al nuevo formato multidimensional. A cambio ofrecen un rendimiento muy superior a la hora de realizar la extracción y el análisis de los datos, puesto que los datos a los que acceden ya están organizados en dimensiones y jerarquías.
- ROLAP (Relational-OLAP): funcionan sobre bases de datos relacionales. Permiten trabajar sobre las bases de datos corporativas ya establecidas, ahorrando así el trabajo de crear y mantener nue-

vas bases de datos multidimensionales. A cambio deben ocuparse de realizar la conversión entre la visión relacional de los datos mantenida por el SGBDR y el manejo multidimensional y jerárquico que debe ofrecer al usuario, lo cual acarrea un coste en tiempo y recursos de máquina.

En ambos casos, las herramientas OLAP, sean del tipo que sean, están especialmente indicadas cuando se quiere obtener una vista dinámica de los datos, estudiar escenarios posibles, comparar unos datos en función de otros sin limitaciones en el tamaño de estas comparaciones (el equivalente al costoso join relacional), obtener datos agregados respecto de series temporales o cualquier otra actividad de análisis de alto nivel.

Respecto a la elección entre MOLAP y ROLAP, en la práctica resulta mucho más habitual encontrar sistemas de almacén de datos, junto con sus correspondientes herramientas OLAP y de minería de datos, implementadas mediante bases de datos relacionales. Esto es debido a la mayor experiencia de que se dispone para trabajar sobre bases de datos relacionales, a la gran cantidad de productos ya disponibles en el mercado y a la confianza que las organizaciones tienen en estos tipos de bases de datos.

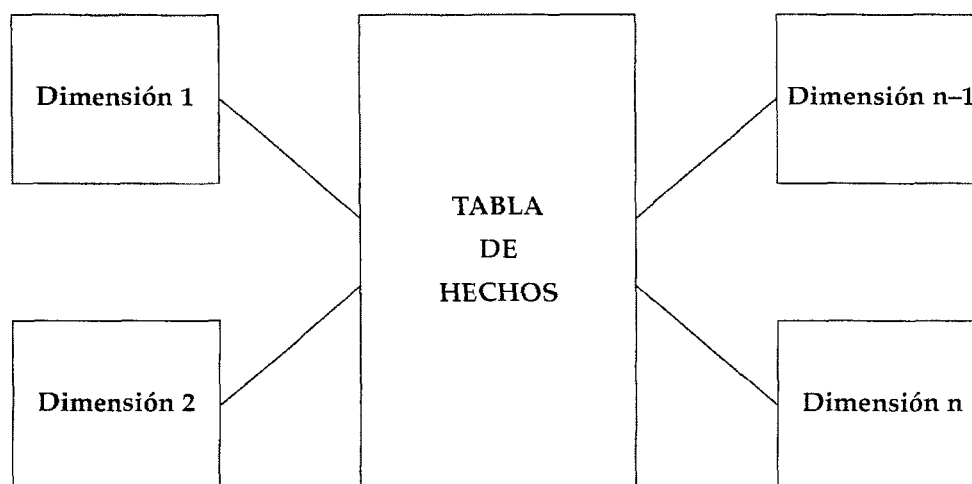
Para simular bases de datos multidimensionales cuando sólo se dispone de un gestor de bases relacional sin que se produzca un efecto negativo en los tiempos de respuesta, se ha extendido la utilización de unos esquemas de datos y unos índices específicamente diseñados a tal efecto. Debido a su importancia práctica vamos a presentar en el siguiente punto estos dos elementos.

9.5. ESQUEMAS EN ESTRELLA E ÍNDICES BITMAP.

9.5.1. Esquema en estrella.

Un esquema en estrella («star schema» en inglés) es una técnica de diseño de bases de datos relacionales que sirve para simular el funcionamiento de bases de datos multidimensionales.

La estructura de las tablas es la que se muestra en la figura siguiente, en la que se puede ver cómo el diseño está compuesto por una única tabla de gran tamaño, llamada tabla de hechos, y por cierto número de tablas mucho más pequeñas que reciben el nombre de tablas de dimensiones.



Un ejemplo práctico de esquema en estrella sería el compuesto por las siguientes relaciones:

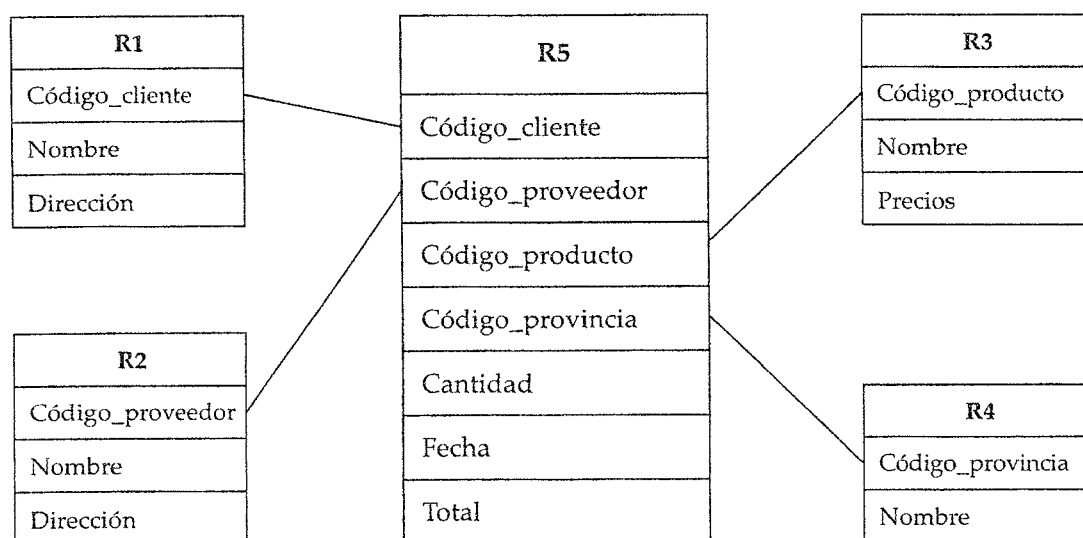
R1: (Código_cliente, Nombre, Dirección)

R2: (Código_proveedor, Nombre, Dirección)

R3: (Código_producto, Nombre, Precio)

R4: (Código_provincia, Nombre)

R5: (Código_cliente, Código_proveedor, Código_producto, Código_provincia, Cantidad, Fecha, Total)

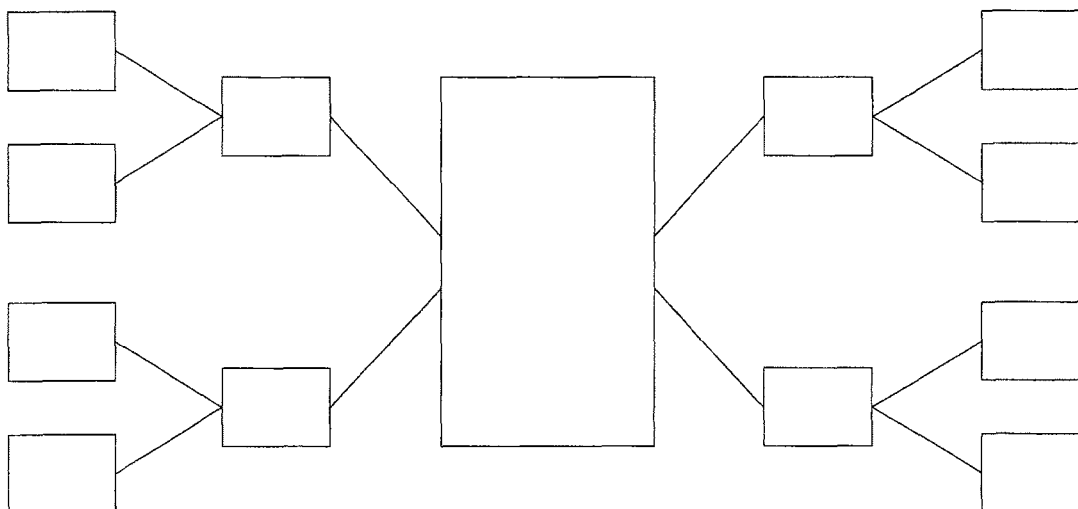


En este ejemplo se ve claramente por qué este tipo de diseño se llama en estrella: existe una tabla central de hechos (R5) que será la que más tuplas contenga, ya que habrá una por cada venta que se haya hecho, y una serie de puntas de la estrella formadas por el resto de tablas (R1-R4). Estas puntas de la estrella se corresponden con las dimensiones que tendría el hipercubo si la base de datos fuera multidimensional. Entre cada tabla de dimensión y la tabla de hechos se establece una relación mediante una clave ajena (también llamada referencial o externa), que va a ser la que se utilice para recuperar cualquier información que se solicite. Las tablas de dimensión no están relacionadas entre sí y sólo se utilizan como puntos de acceso a los datos detallados de la tabla de hechos.

Al igual que sucede al manejar un hipercubo multidimensional, las consultas típicas en un esquema en estrella consisten en fijar un valor o un rango de ellos para las dimensiones y, a continuación, obtener la información solicitada. La respuesta se encuentra realizando operaciones de unión natural (join) entre las tablas de dimensión y la de hechos.

Para optimizar estas consultas, el gestor de bases de datos debe ser capaz de reconocer que está trabajando con un esquema en estrella y hacer en primer lugar los join entre las tablas de dimensiones y, con el resultado, hacer un único join con la tabla de hechos, minimizando así el número de accesos físicos.

Otra estructura empleada en modelos relacionales para simular modelos multidimensionales es el llamado esquema en copo de nieve («snowflake schema» en inglés). Se utiliza cuando hay jerarquías en las dimensiones y más claves ajenas. Su estructura es la siguiente:



Podríamos utilizar un modelo de copo de nieve en el problema de nuestro ejemplo si para cada producto (R3) almacenáramos además su fabricante o, si dentro de cada provincia (r4) distinguiéramos zonas de venta. En estos casos hay que crear relaciones adicionales con sus correspondientes claves ajenas. Además ahora las claves ajenas no están sólo en la tabla de hechos, sino también las de dimensiones.

9.5.2. Índices bitmap.

Los índices bitmap son un tipo especial de índice que almacena la información en bits en vez de múltiplos de bit (byte, doble byte, etc.) y que sirve para acelerar el acceso a tuplas con atributos de baja cardinalidad.

Se dice que un atributo es de baja cardinalidad si su dominio está formado por pocos elementos. Por ejemplo, el atributo «sexo» es de baja cardinalidad porque sólo puede tomar dos valores («H» o «M»). En cambio, el atributo DNI no es de baja cardinalidad, porque puede tomar millones de valores distintos.

Dada la siguiente tabla:

DNI	NOMBRE	APELLIDOS	ESTADO CIVIL	SEXO	EDAD
1234567890	Juan	Rubio Sanz	S	M	32
1112456090	Ana	García Martín	C	F	22
0923456948	Rosa	Pérez Alonso	V	F	43

Se puede definir un índice bitmap para el atributo «sexo» y otro para el atributo «estado civil». El índice para «sexo» está compuesto por dos bitmaps, uno para cada posible valor del atributo, tal como se muestra a continuación:

SEXO = H	SEXO = M
1	0
0	1
0	1

Este índice nos indica que la primera tupla de la tabla tiene «H» como valor del atributo «sexo», mientras que las otras dos tuplas tienen el valor «M». Aquí se ve que hay que guardar un bitmap para cada posible valor del atributo, por lo que, como se dijo anteriormente, no es eficiente usar estos índices para valores de alta cardinalidad.

Igualmente, el índice para «estado civil» sería el siguiente:

ESTADO CIVIL = S	ESTADO CIVIL = C	ESTADO CIVIL = V	ESTADO CIVIL = D
1	0	0	0
0	1	0	0
0	0	1	0

Para responder a consultas que se realicen sobre esquemas relacionales con índices bitmap, basta con hacer las operaciones lógicas apropiadas (AND, OR, NOT) sobre los bits de cada índice implicado en la consulta, lo cual es una operación muy rápida, mucho más que la comparación de cadenas o números que implica la utilización de índices de otro tipo.

Este tipo de índices son útiles para indexar las tablas de dimensiones en esquemas en estrella o en copo de nieve, ya que muchas de estas dimensiones suelen tener su clave principal formada por un atributo de baja cardinalidad (p. ej. código de provincia, sexo, estado civil, mes, año, comunidad autónoma, etc.).

Una vez vistas estas consideraciones prácticas sobre la implementación de modelos multidimensionales sobre bases de datos relacionales, vamos a pasar a ver los criterios de elección de una herramienta OLAP.

10. ELECCIÓN DE UNA HERRAMIENTA OLAP.

A la hora de elegir una herramienta OLAP hay que tener en cuenta, entre otros, los siguientes puntos:

- Si obligan a trabajar con una base de datos multidimensional (MOLAP), relacional (ROLAP) o si soportan ambas.
- En el caso de herramientas MOLAP es conveniente estudiar las capacidades de la BDM subyacente. Además hay que fijarse en su capacidad de aceptar accesos concurrentes y la carga de usuarios que admiten, ya que si el objetivo del OLAP es permitir el análisis interactivo estos dos puntos van a determinar en gran medida la utilidad de la herramienta.
- En el caso de herramientas ROLAP la penalización en que se incurre al no utilizar una base de datos multidimensional y las facilidades que ofrece la herramienta para ofrecer una vista multidimensional de los datos relacionales (optimización de accesos a esquemas en estrella, en copo de nieve e índices bitmap).
- El límite en cuanto al número de dimensiones y de celdillas que puede manejar, sea o no multidimensional la base de datos subyacente. También la profundidad de los niveles de jerarquías y el manejo de series temporales.
- La capacidad de cálculo y la facilidad para especificar qué métodos y operaciones hay que aplicar a los datos. También debe disponer de herramientas de formateo y presentación de informes.
- El mantenimiento de las dimensiones y las jerarquías mediante herramientas automatizadas. Facilidad a la hora de modificar cualquiera de ambos elementos.

Además, antes de lanzarse a implementar estas tecnologías (almacén de datos, minería de datos, OLAP) hay que tener muy claro el coste que esto va a suponer. Algunos de los elementos a considerar son los siguientes:

- Esfuerzo necesario para determinar el estado de los datos disponibles a partir de los sistemas OLTP y realizar su limpieza y migración.
- Establecimiento de la estructura interna del almacén de datos así como del SGBD que se vaya a utilizar.
- Elección y adquisición de los equipos y el software necesarios para soportar tanto el almacén como las herramientas de minería.
- Dificultad en encontrar expertos en la materia que asesoren en el proceso de construcción del sistema de gestión de datos.
- Necesidad de obtener un compromiso por parte de los altos niveles directivos para facilitar el acceso a los datos pertenecientes a las diversas unidades dentro de la organización.
- Mantenimiento y actualización continua de las bases de datos que componen el almacén. En el caso de utilización de BDM, aprendizaje de lo que probablemente será a una nueva forma de almacenar y acceder a la información.

BIBLIOGRAFÍA

- Temario de las pruebas selectivas para ingreso en el Cuerpo Superior de Sistemas y Tecnologías de la Información de la Administración del Estado. ASTIC.
- Temario de las pruebas selectivas para el acceso, por promoción interna, al Cuerpo de Gestión de Sistemas e Informática de la Administración del Estado. MINISTERIO PARA LAS ADMINISTRACIONES PÚBLICAS.
- Data Mining Notes. http://www-pcc-ac.uk/tec/cours.../stu_notes/dm_book_2.html
- OLAP Council White Paper. <http://www.olapcouncil.org/research/whtpapco.htm>
- *Analysis of Data Mining Algorithms*. KARUNA PANDE JOSHI
- *IMS Learning Design Information Model*. IMS Global Learning Consortium, Inc.

