

Capítulo 21. NUEVAS APLICACIONES

1. SISTEMAS DE AYUDA A LAS DECISIONES

Partiendo de una base de datos con información sobre clientes, ventas y otros datos, una empresa puede utilizar esa información para identificar las necesidades del mercado y anticiparse a la demanda, o centrar sus esfuerzos en un determinado camino.

- Es necesario el almacenamiento y la recuperación de los datos para la ayuda de decisiones, pero esto genera varios problemas.
 - Aunque muchas de las consultas pueden realizarse directamente en SQL, otras no se pueden o no resulta fácil
 - Los lenguajes de bases de datos no son adecuados para análisis estadísticos
 - Las técnicas de *inteligencia artificial* intentan descubrir las pautas de los datos, pero hay que combinarlas con una implementación eficiente de grandes bases de datos.
 - Las grandes empresas reciben datos de diferentes fuentes geográficas, y se almacenan de manera que la recuperación de estos por parte de otros departamentos no es rápida.

2. ANÁLISIS DE LOS DATOS

Los datos guardados en una base de datos suelen ser de gran tamaño, y hay que resumirlos de alguna manera, SQL incorpora algunas funciones de agregación, pero no son suficientes y la mayoría de las bases de datos incluyen un conjunto más amplio (*varianza*, *media*, etc).

También son utilizadas en análisis estadístico las *tablas cruzadas*:

	Talla pequeña	Talla mediana	Talla grande	Total
Color claro	8	35	10	53
Color oscuro	20	10	5	35
Total	28	45	15	88

Una tabla cruzada es multidimensional (bidimensional en el ejemplo), no se corresponde con una tabla relacional y no puede generarse con una sola consulta SQL.

3. RECOPIACIÓN DE LOS DATOS

Se trata de buscar información relevante entre todos los datos almacenados al igual que la *inteligencia artificial*, la recopilación de datos intenta conocer de manera automática las reglas estadísticas y las pautas a partir de los datos grabados (normalmente en discos).

Representación de la información mediante reglas → La información obtenida mediante la recopilación de datos debe representarse mediante un conjunto de reglas que proporcionan una infraestructura común, las reglas son de la forma:

$\forall X \text{ antecedente} \Rightarrow \text{consecuente}$ [Ejemplo: $\forall T, compra(T, pan) \Rightarrow compra(T, leche)$]

Donde T es una variable cuyo rango son todas las transacciones.

- **Soporte:** es una medida de la parte de población que cumple la regla, las reglas que tienen poco soporte no son significativas.
- **Confianza:** es la medida de la frecuencia con que el consecuente se cumple al cumplirse el antecedente (Ejemplo. Si el 80% de las personas que compran pan, también compran leche, entonces la confianza es del 80%). Al igual que con el soporte, la reglas con confianza baja no son significativas.

Clases de problemas de recopilación de datos → La Clasificación y las Reglas de Asociación; que implica encontrar reglas que determinen conclusiones a partir de los datos. (Ejemplo: puede que todo el que compre pan compre leche, o puede que no sea así, hay que clasificar y definir reglas que sean significativas)

Recopilación de datos dirigida por el usuario → El usuario tiene la responsabilidad principal en el descubrimiento de las reglas, la base de datos representa un papel secundario.

Descubrimiento automático de las reglas → Se ha visto beneficiado por los avances en *inteligencia artificial*, con la diferencia del volumen de datos a manejar y que requiere de algoritmos especializados en el manejo de datos almacenados en disco.

4. ALMACENES DE DATOS

Un almacén de datos es un depósito de la información reunida a partir de varias fuentes, guardada según un esquema unificado en un único lugar, estos datos se guardan durante mucho tiempo para permitir el acceso a datos históricos, los almacenes de datos proporcionan al usuario que debe tomar las decisiones una interfaz única para todos los datos, además de no cargar de trabajo a las bases de datos para el procesamiento de transacciones.

Momento y manera de recoger los datos → Pueden transmitirse los datos a medida que se realizan las transacciones o bien transmitirlos periódicamente, el almacén suele tener algunos datos sin actualizar.

Esquema que debe utilizarse → Puesto que los orígenes de datos son independientes puede que utilicen modelos diferentes de datos, así pues habrá que integrar los diferentes tipos en el almacén.

Propagación de las actualizaciones → Hay que realizar un mantenimiento de las relaciones que se modifiquen en los orígenes de datos respecto al almacén.

Datos que se deben resumir → Los datos pueden agruparse siempre que no represente una merma en la capacidad para la toma de decisiones.

5. BASES DE DATOS GEOGRÁFICAS Y ESPACIALES

Estas bases de datos guardan información relacionada con ubicaciones espaciales: Bases de Datos para Diseño (CAD); Bases de Datos Geográficas (Mapas)

6. BASES DE DATOS MULTIMEDIA

Son bases de datos que guardan imágenes, sonido y vídeo, estos datos no se almacenan en la base de datos, sino fuera, en el sistema de archivos, pero esto puede provocar inconsistencia, puede que un archivo figure en la base de datos y sin embargo se halla borrado, o viceversa. Es preferible guardar los datos dentro de la propia base de datos.

Recuperación Basada en la Semejanza → Los índices no pueden crearse igual que en bases de datos tradicionales, pues puede hacer falta buscar datos por aproximación, que requieren de índices especiales, pueden buscarse por semejanza: Imágenes, Sonidos y Manuscritos

Datos de Medios Continuos → Son tipos de datos de medios continuos el vídeo y el sonido, que requieren ser suministrados a una velocidad y un ritmo constante y sincronizado. Estos archivos suelen estar en formato comprimido como JPEG y MPEG.

7. COMPUTADORAS PORTÁTILES Y BASES DE DATOS PERSONALES

Modelo para computadoras portátiles → Consiste en computadoras portátiles y una red de computadoras unidas mediante cables, los portátiles se conectan con la red mediante *estaciones de apoyo para computadoras portátiles*. Donde cada estación administra las computadoras portátiles de su área geográfica o *celda*.

Encaminamiento y procesamiento de consultas → El modelo para computadoras portátiles implica que la ruta entre dos de ellas puede variar, esto supone que las direcciones de la red basadas en la ubicación ya no son constantes.

Difusión de datos → Es conveniente que las *estaciones de apoyo* transmitan los datos solicitados con mayor frecuencia en un ciclo continuo, en lugar de a petición de las computadoras portátiles, motivos:

- Ahorro energético de la computadora portátil al no tener que solicitarla.
- La *estación de apoyo* puede transmitir la misma información a muchas portátiles a la vez.

Desconexiones y consistencia → Puesto que las conexiones suelen pagarse por tiempo, es deseable desconectar algunas computadoras portátiles durante periodos largos, sin embargo el usuario puede seguir ejecutando consultas y actualizaciones desde el portátil, problemas:

- *Recuperabilidad*: Las actualizaciones introducidas en el portátil pueden perderse si la computadora sufre un fallo.
- *Consistencia*: Los datos almacenados en el portátil pueden quedarse desfasados entre una conexión y otra.
- *Actualización de datos en distintas computadoras desconectadas*: Este problema se resuelve mediante el intercambio de *vectores de versión* entre las computadoras afectadas.

Debido a estos problemas resulta preferible que los usuarios preparen las transacciones y las transmitan para su ejecución en el servidor

8. SISTEMAS DE RECUPERACIÓN DE LA INFORMACIÓN

El proceso de recuperación de la información consiste en localizar los documentos de importancia (catálogos de bibliotecas, sistemas de gestión de documentos, etc.)

Consultas → Se trata de la recuperación de la información mediante palabras clave introducidas por el usuario.

Creación de índices de documentos → Cada palabra clave puede estar en un gran número de documentos, luego es fundamental una representación compacta para que el índice haga poco uso del disco

Exploración e hipertexto → Los sistemas de hipertexto se concibieron inicialmente como sistemas para catálogos, y tienden a utilizar una jerarquía de clasificación (como la de una biblioteca), donde un documento puede estar incluido en varias categorías, donde habrán punteros al documento real.

9. SISTEMAS DE INFORMACIÓN DISTRIBUIDOS

Con la aparición de Internet se han desarrollado varios sistemas de información distribuida

- *Gopher*: Existen servidores y clientes, los servidores *Gopher* organizan los datos en directorios, los clientes se conectan al servidor y muestran los directorios de nivel superior en forma de menú, cada opción de menú puede ser un documento, otro directorio o un enlace a otro servidor *Gopher*.
- *Sistema de información de área amplia WAIS*: Crea índices de información por los sistemas de servidores, cada servidor guarda información sobre uno o varios temas y también sobre que temas guardan otros servidores.

10. WORLD WIDE WEB

Es un sistema de información distribuido basado en el hipertexto, el formato de los documentos de hipertexto es el HTML, mostrando texto con formato e imágenes, mucho más atractivo que un simple menú.

- *Localizadores universales de recursos (URL)*: Son direcciones únicas de un documento en toda Internet, la primera parte indica el protocolo (http, ftp) la segunda indica la dirección de la máquina (www.dell.com) y la tercera indica la dirección del documento dentro de la máquina (/dir/set/book).
- *Servidores Web*: Debido a la potencia del protocolo http, los servidores web sirven como interfaz para gran variedad de servicios de información.
- *Lenguajes de visualización*: Los lenguajes de texto con formato tradicionales eran lentos e inadecuados para el uso interactivo.
 - *Lenguaje de marcas de hipertexto*: SGML proporciona una gramática para los formatos de los documentos basada en anotaciones de marcas normalizadas, HTML está basado en él. Por lo tanto HTML proporciona una interfaz gráfica de usuario que puede ejecutarse en casi todas las computadoras.
 - *Java*: El lenguaje Java permite que los documentos sean activos, permitiendo ejecutar programas en la computadora cliente, el código compilado de Java necesita ser interpretado, actualmente existen intérpretes Java para cualquier soporte y sistema operativo.
- *Interfaces Web para las bases de datos*: El Web con el HTML forman una interfaz adecuada para la creación de transacciones, donde se rellena un formulario, se pulsa un botón y el servidor ejecuta la operación, devolviendo al usuario el resultado en el mismo HTML. El problema reside en la actualización de los documentos HTML, que quedan desfasados rápidamente, es necesario la creación de documentos de manera dinámica. Las herramientas que convierten SQL en HTML, para mostrar al usuario los datos de la consulta son parecidas a los *generadores de informes*.
- *Búsqueda de información en el sistema Web*: El enorme crecimiento de la información guardada en el sistema Web representa un problema a la hora de encontrar lo que resulta interesante. Se han desarrollado varios sistemas de recogida de información denominados *Web crawlers*, estos, siguen de manera recursiva los enlaces de páginas conocidas hacia otras desconocidas, creando índices. La consulta de estos documentos se realiza con la técnica de las *palabras clave*.