Sistemas de soporte a las decisiones

Primer parcial

Ultima modificación el lunes 26 de noviembre de 2007 a las 17:52:25

Business intelligence

Conceptos y metodologías para mejorar la toma de decisiones en los negocios basándose en hechos y sistemas que trabajan con hechos.

Recursos y herramientas

- Fuentes de datos : datawarehouse, datamarts, etc
- Herramientas de administración de datos
- Herramientas de extracción y consulta
- Herramientas de modelización (Data mining)

Evolución de los querys hacia la minería de datos

1. Querys de Data lleva a Observar
2. Análisis de información lleva a Analizar
3. Data mining de conocimiento lleva a Predecir/Pronosticar

Analista de datos

Vincula área de tecnología informática con área de negocios.

Traduce los requerimientos de información en preguntas apropiadas para su análisis con las herramientas de minería.

Realimenta el Data Warehouse de la compañía con nuevos criterios de data cleaning y data validation.

Habilidades requeridas

- Manipulación de datos (SQL)
- Conocimiento de las técnicas de minería y análisis exploratorio
- Habilidad de comunicación e interpretación de los problemas del negocio
- Creatividad

Data warehouse

Colección de datos, integrada, no volátil, varia en el tiempo y organizada a partir de un tema para soportar decisiones de negocios.

Es una copia de los datos transaccionales específicamente orientada para querys y análisis.

Componentes: fuentes de datos, extracción, transformación y carga ETL, soporte físico de los datos DBMS, herramientas de exploración.

Data marts

Técnicamente es un subconjunto del Data Warehouse orientado a una finalidad especifica del negocio : marketing, finanzas, producción, etc.

También se usa para identificar soluciones alternativas al DW, mas reducidas y de menor costo y menor tiempo de implementación.

Herramientas de explotación del Data Warehouse

Herramientas de visualización

Reporting

OLAP

On line Analytical Processing, permite la elaboración de vistas multidimensionales del DW para optimizar performance. Motores de administración del DW permiten la construcción de estos “Cubos”.

Complementa el data mining, y supera las posibilidades de SQL.

Data Mining

Es un proceso para la exploración y análisis de manera automática o semiautomática de los datos para obtener patrones significativos y reglas de negocio.

Los patrones obtenidos deben ser significativos.

No es un producto que se compra, es una disciplina a dominar.

No es una solución instantánea de los problemas del negocio.

No es un fin en si mismo, sino un proceso que ayuda a encontrar soluciones a problemas de negocio.

Sus pilares son

- Datos
- Algoritmos y tecnicas
- Practicas de modelización

Integra las disciplinas

- Inteligencia artificial
- Estadística
- Tecnología de soporte de decisiones OLTP
- Tecnologías de hardware y software

Etapas en el proceso de data mining

- Identificar el problema del negocio
- Transformar los datos en información
- Actuar a partir de los resultados
- Medir los resultados de las acciones.

Objetivos principales

- Predicción
- Descripción

Pasos del proyecto genérico

- Entender el negocio
- Entender los datos
- Preparar los datos
- Modelar
- Evaluar
- Implementar
- Feedback (volver al principio)

Data mining y OLAP

Las herramientas de reporting, OLAP y consulta permiten construir modelos descriptivos y retrospectivos, para confirmar o rechazar hipótesis previas del usuario.

Las herramientas de data mining permiten encontrar patrones no evidentes en los grandes volúmenes de información del DW y proponer modelos predictivos.

Estadística y minería de datos

La estadística es la disciplina que extrae información general a partir de datos específicos. Estudia la estabilidad de la variación. Examina, sumariza y extrae conclusiones a partir de los datos.

Los métodos estadísticos son el corazón del data mining.

En la minería de datos no se hacen supuestos a priori de la naturaleza de las variables y las relaciones entre ellas.

Los algoritmos estadísticos se adaptan al procesamiento de grandes volúmenes de datos.

IA y minería de datos

La IA se integra a partir de las redes neuronales, se usa para construir modelos predictivos no lineales que aprenden a traves del entrenamiento.

Las redes neuronales son adecuadas para problemas de tipo predictivos.

Customer Relationship Management

Proceso que administra la relación entre la empresa y los clientes.

Es necesario para el éxito identificar los patrones de consumo y comportamiento de los clientes.

Medidas de efectividad

Retorno de la inversión, y tasa de error entre realidad y modelo predictivos.

Proyecto Exitoso

Un único project leader, equipo multi disciplinario, involucrar todas las áreas desde el comienzo, y comenzar con un pequeño proyecto piloto de data mining.

Análisis exploratorio de datos

Abreviado AED o DEA, sirve para maximizar la comprensión del conjunto de datos, extraer las variables importantes, detectar anomalías, y probar supuestos necesarios (linealidad, normalidad, etc).

Etapas del AED

- Acceso y Preparación de los datos
- Descripción univariada (numérica y gráfica)
- Descripción multi variada (numérica y gráfica) para identificar relaciones subyacentes básicas.
- Evaluación de supuestos estadísticos (normalidad, linealidad, etc)
- Identificar outliers (casos atípicos)
- Evaluación del impacto de los casos perdidos (missing values)

Escala de medida de variables

Escala	**Gráficos**	**Tendencia central**	**Dispersión**
Nominal	Barras-Lineas-Sectores	Moda	
Ordinal	Boxplot	Mediana	Rango intercuartil
Intervalos	Histograma – Polígono de frecuencias	Media aritmética	Desviación típica
Ratios		Media geométrica	Coeficiente de variación

Análisis estadístico bidimensional

Posibilidades

1. ambas variables medidas en escalas nominales u ordinales (no métricas)
2. ambas variables medidas en escalas de ratios o proporciones (métricas)
3. una variable medida en escala nominal-ordinal y otra en escala de proporciones.

Metodología

Variables	Metodología
ambas variables medidas en escalas nominales u ordinales (no métricas)	Tabla de contingencia
	Prueba de Chi2
	Perfiles de filas
ambas variables medidas en escalas de ratios o proporciones (métricas)	Diagrama de dispersión
	Coeficiente de correlación
	Recta de regresión
una variable medida en escala nominal-ordinal y otra en escala de proporciones.	Diagrama de cajas boxplot
	Error bar
	Prueba t comparación de medias

Metadata

Describe los datos del datawarehouse. Vincula los datos con los usuarios del negocio.

Información sobre los datos tal como fuente de datos, descripcion de transformaciones, estructura de datos del DW, reglas de clean up, referencias historicas y temporales, etc...

ETL

Extracción, transformación (limpieza) y carga (load) de los datos en el datawarehouse.

Adquisición y limpieza de los datos

Los objetivos son remover datos no necesarios de las fuentes operacionales, consolidar representaciones de datos de diferentes fuentes, calcular sumarizaciones y variables derivadas, y resolver problemas de missings y outliers.

Integridad de datos

Los datos cumplen condiciones de integridad cuando se ajustan a todos los estándares de valor y completitud. Todos los datos del DW son correctos, y el DW esta completo (no existen mas datos fuera de él).

La **credibilidad** del DW depende de la integridad de sus datos.

Modelo dimensional

Hechos

Los hechos se registran en las tablas centrales del DW. Representan un item de negocio, una transacción o un evento.

Cada hecho tiene dimensiones asociadas.

Dimensiones

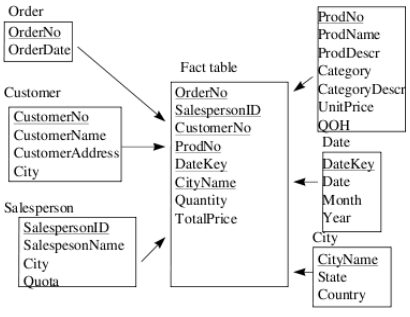
Una dimensión es una colección de miembros o unidades o individuos del mismo tipo. Cada punto de entrada de la tabla de **hechos** esta conectada a una dimensión.

Determinan el contexto de los hechos.

Algunas dimensiones habituales son tiempo, geografía, cliente, vendedor, etc...

Modelos básicos dimensionales

Estrella



Copo de nieve

Modelo ER vs Modelo Dimensional

El modelo ER (entidad relación) utiliza foreign keys y entidades.

Las foreign key serian las dimensiones, y la entidad el hecho.