

Capitolo 12° Data mining e Warehousing spaziali

Data mining e Warehousing spaziali

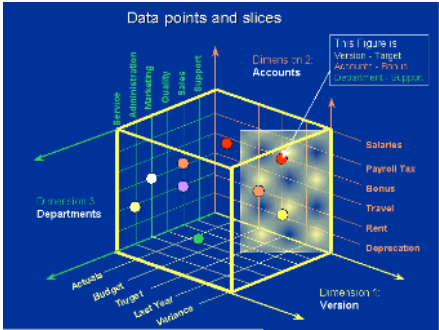
- 12.1 – Riutilizzo dei vecchi dati
- 12.2 – Datawarehousing
- 12.3 – Data mining
- 12.4 – Data mining spaziale
- 12.5 – Conclusioni

12.1 – Riutilizzo dei vecchi dati

- Conservare i dati per motivi giuridici dando una struttura comune che consente di fare delle ricerche
 - ➔ Datawarehouses (emporio di dati)
- Estrarre il succo dei dati
 - ➔ Data mining

12.2 – Datawarehousing

- Archivi storici
- Emporio di dati
- Sistemi OLAP
- Cubi



OLAP = Online Analytical Process

What is a Warehouse?

- Collection of tools
 - ◆ gathering data
 - ◆ cleansing, integrating, ...
 - ◆ querying, reporting, analysis
 - ◆ data mining
 - ◆ monitoring, administering warehouse

CS 245

Notes12

5

Warehouse Architecture

```
graph TD; Client1([Client]) --> QA[Query & Analysis]; Client2([Client]) --> QA; QA --> Warehouse[(Warehouse)]; Metadata[Metadata] --- Warehouse; Warehouse --> Integration[Integration]; Integration --> Source1[(Source)]; Integration --> Source2[(Source)]; Integration --> Source3[(Source)];
```

CS 245

Notes12

6

OLTP vs. OLAP

OLTP	OLAP
• Mostly updates	• Mostly reads
• Many small transactions	• Queries long, complex
• Mb-Tb of data	• Gb-Tb of data
• Raw data	• Summarized, consolidated data
• Clerical users	• Decision-makers, analysts as users
• Up-to-date data	
• Consistency, recoverability critical	

CS 245

Notes12

13

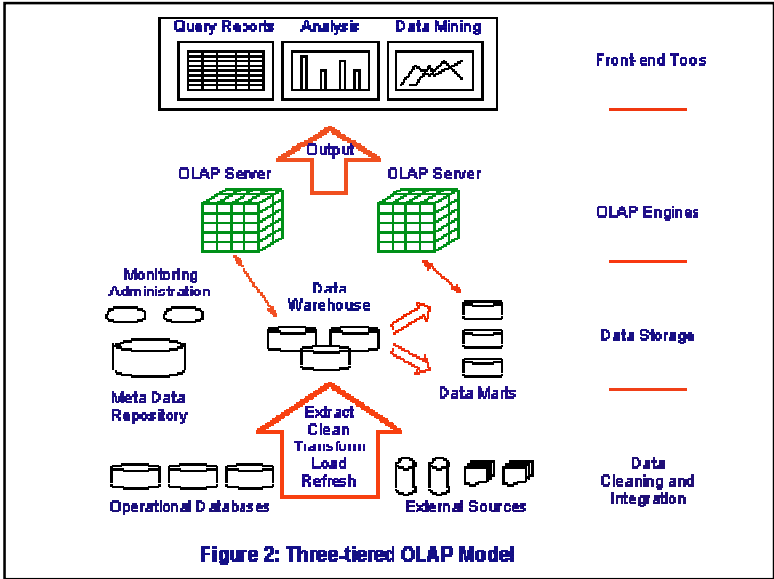
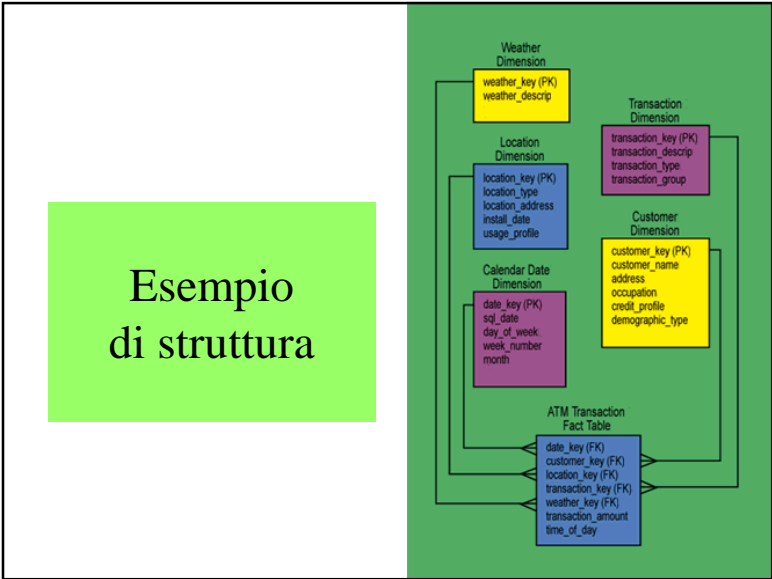
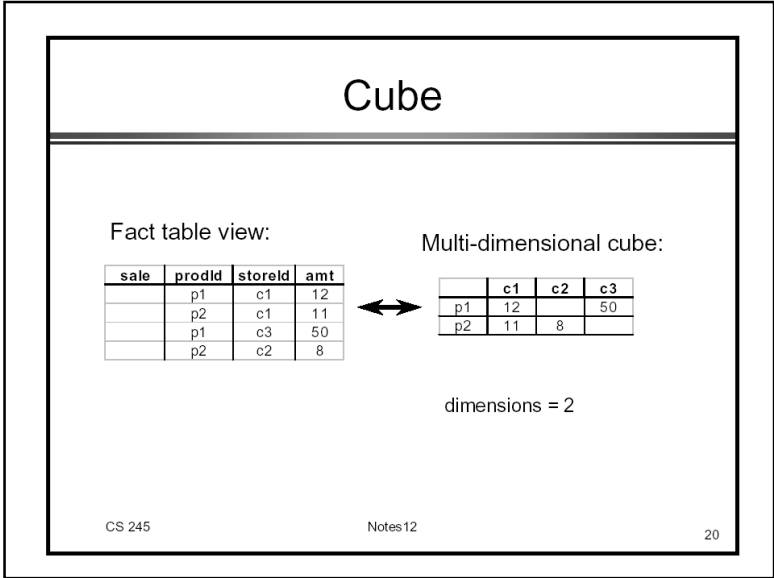
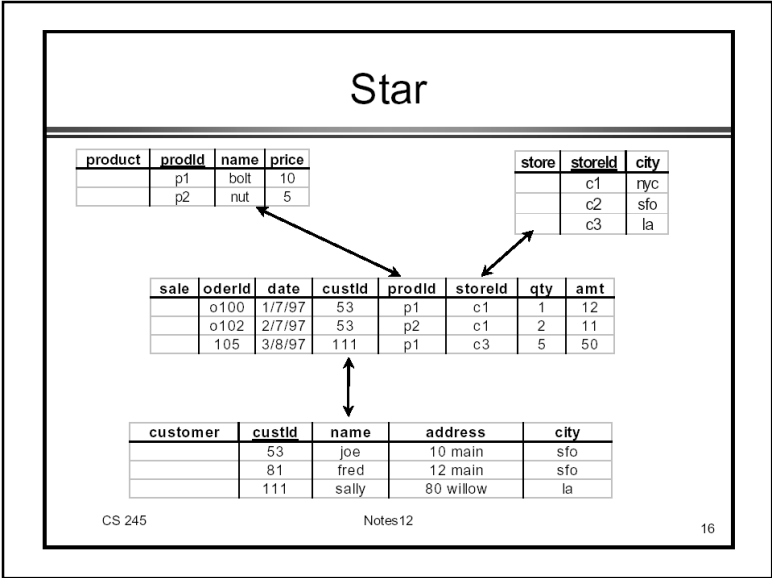
Warehouse Models & Operators

- Data Models
 - ◆ relations
 - ◆ stars & snowflakes
 - ◆ cubes
- Operators
 - ◆ slice & dice
 - ◆ roll-up, drill down
 - ◆ pivoting
 - ◆ other

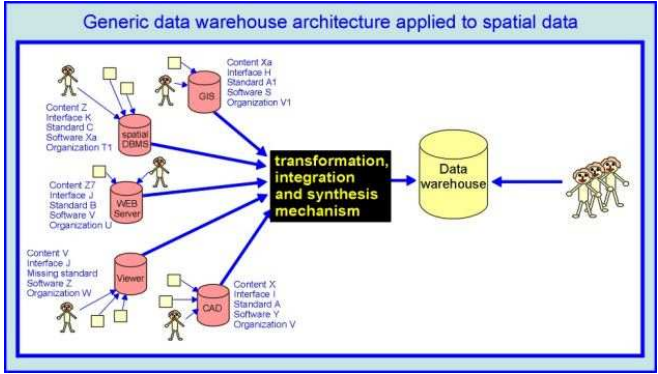
CS 245

Notes12

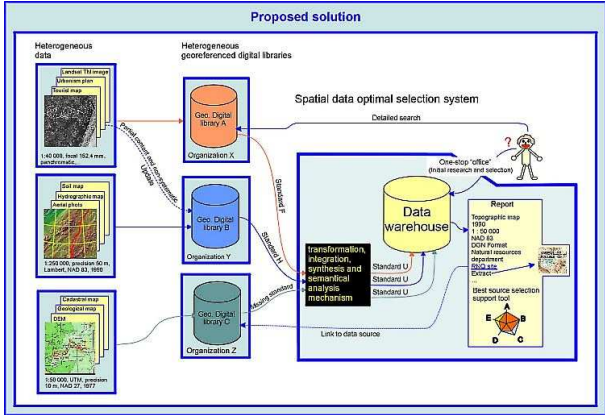
15



Datawarehouse spaziale



Datawarehouse spaziale



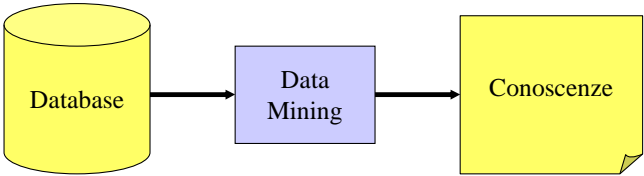
12.3 – Data mining

- Abbiamo tanti dati
- Estrarre il "succo" dei dati
- Metafora:



Database = cava aurifera
Conoscenza = pepite d'oro

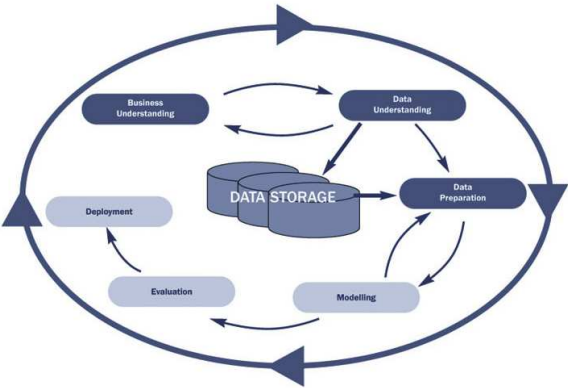
Processo



"Cesto della casalinga"

- Esempio DB relazionale
- *Compere* (*n° compera, pane, latte, burro, marmellata, insalata, carne*)
- Verificare : *pane e latte* → *burro*
- Supporto
- Percentuale di verità

Metodologia generale del Data Mining



12.4 – Data mining spaziale

- Applicazione del data mining agli aspetti spaziali
- Trovare regolarità che possono essere utili nella pianificazione del territorio

Esempio

- Relazione tra
 - lago
 - strada
 - ristorante



Conoscenza del territorio

- Criminalità
- Incidenti stradali
- Uso del suolo
- Inquinamento
- Economia
- Demografia
- ecc.

Esempio criminalità

2000 Incidents of Crime by Census Tract in San Antonio

Source: San Antonio Police Dept.

Mappe di criminalità a Tokyo

2002 Tokyo Crime Map

Tokyo Metropolitan Police

Scales Indicate Incidents per square kilometre

Fasi principali del Data Mining

```
graph LR; A[Estrazione dei dati] --> B[Preparazione dei dati]; B --> C[Analisi dei dati]
```

Prima legge di Tobler

- *"Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970)"*
- Allora !!!
- Scopo: trovare altre leggi

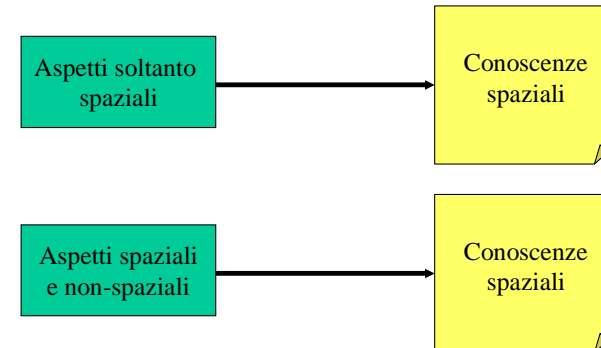
Approcci del Data Mining Spaziale

- Generalizzazione
 - Gerarchia di concetti
- Clustering.
 - Processo di raggruppamento d'oggetti fisici o astratti in classi d'oggetti simili
 - Approcci: Partitioning, Partitioning gerarchico, località, griglie.
- Associazioni spaziali.
 - Regole che descrivano le implicazioni tra diversi caratteristiche, verso un altro gruppo di caratteristiche.

Approcci del Data Mining Spaziale

- Approssimazione ed Aggregazione
 - Perché abbiamo lì una clusterizzazione ?
- Mining in database d'immagini o di raster
 - Può essere visto come un caso particolare.
- Classificazione spaziale
 - Assegnazione di un oggetto a una classe basata sui valori degli attributi
- Scoperta di un trend spaziale
 - Evoluzione regolare di un insieme d'attributi nell'intorno di un altro oggetto

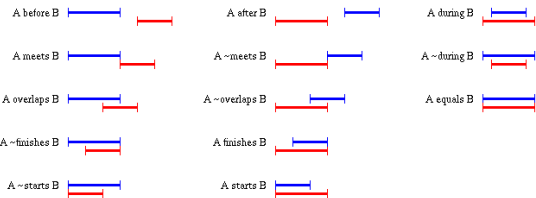
Idee generali



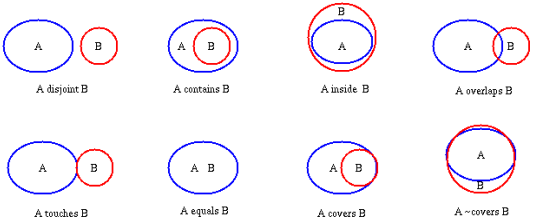
Contesto

- Esistono algoritmi di data mining basati sui grafi
- Rappresentazione grafica
 - dati non-spaziali
 - dati spaziali
 - relazioni spaziali e non-spaziali
- Relazioni spaziali
 - topologiche
 - direzionali
 - distanza
 - mereologiche

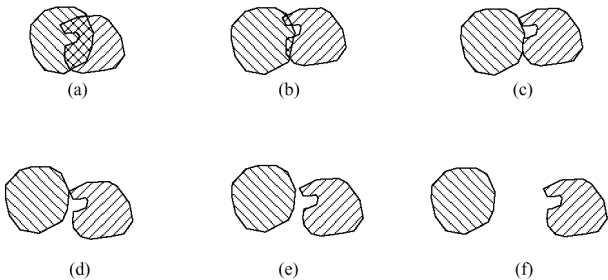
Allen



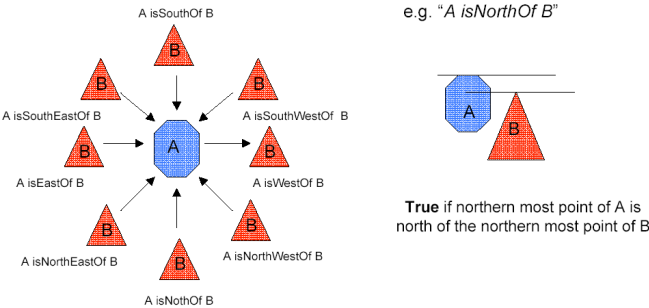
Egenhofer



Esempi di relazioni topologiche



Relazioni direzionali



Relazioni di distanza

Quantitative Distance

A

← 5 km →

B

Qualitative Distance

A is close to B

A

B

meters

Relazioni mereologiche

Hawaii isPartOf United States

Esempio: incidenti stradali

Incidenti

Rete viaria

Comuni

...

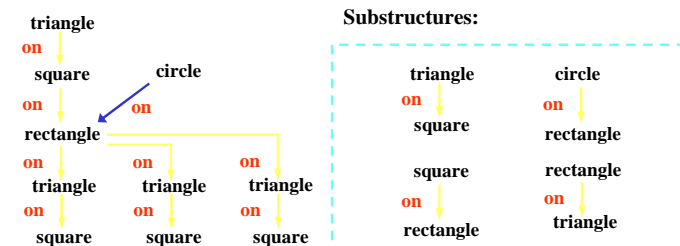
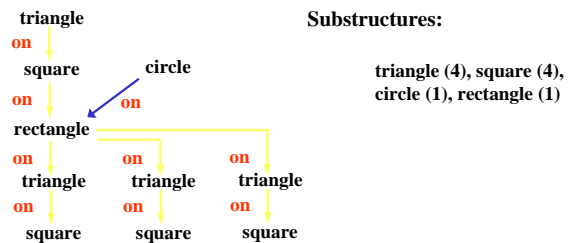
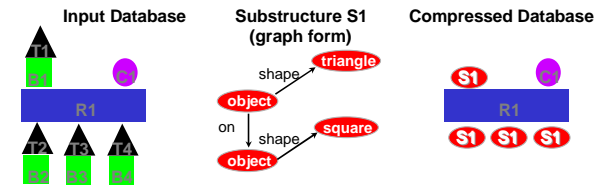
Esempio:
caratterizzare gli incidenti notturni

Generalizzazione della tabella
« incidenti » su due attributi:
giorno e luminosità

Ricerca delle concentrazioni
spaziali locali anormali

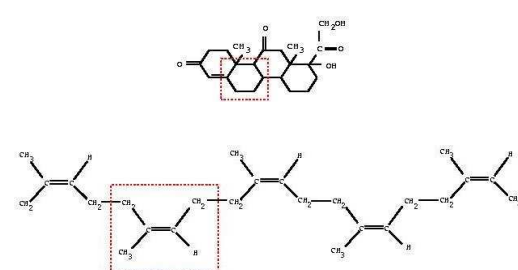
- Trasformazione dei dati in un grafo
 - vertici: oggetti ed attributi
 - spigoli: relazioni tra di loro
- Uso di un algoritmo di ricerca nei grafi
- Non necessario mettere tutti gli oggetti e tutti i loro attributi nel grafo
- Ricerca per pattern particolari
- Uso dell'algoritmo SUBDUE

- Sviluppato all'Università del Texas
- Ricerca nei grafi
- <http://cygnus.uta.edu/subdue>



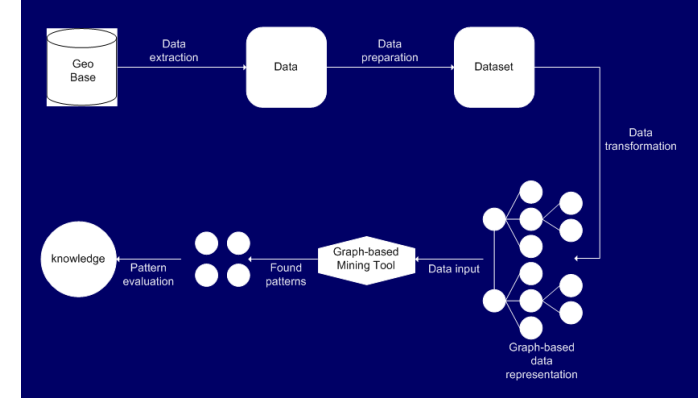
Ricerca della ripetizione di una sotto-struttura

Esempi



The image shows two chemical structures. The top structure is a complex polycyclic molecule with a central ring system and various substituents, including a carboxylic acid group. A red dashed box highlights a specific sub-structure within it. The bottom structure is a long, branched hydrocarbon chain with several double bonds and methyl groups. A red dashed box highlights a specific sub-structure within it.

Approccio di Pech



The flowchart illustrates the Pech approach. It starts with a 'Geo Base' (represented by a cylinder) leading to 'Data' (a rounded rectangle) via 'Data extraction'. 'Data' then leads to 'Dataset' (a rounded rectangle) via 'Data preparation'. 'Dataset' leads to 'Graph-based data representation' (a graph structure) via 'Data transformation'. 'Graph-based data representation' leads to 'Data input' (a rounded rectangle) via 'Data input'. 'Data input' leads to 'Found patterns' (a rounded rectangle) via 'Pattern evaluation'. 'Found patterns' leads to 'Knowledge' (a circle) via 'Found patterns'. 'Knowledge' leads to 'Graph-based Mining Tool' (a rounded rectangle) via 'Found patterns'. 'Graph-based Mining Tool' leads to 'Data input' via 'Data input'.

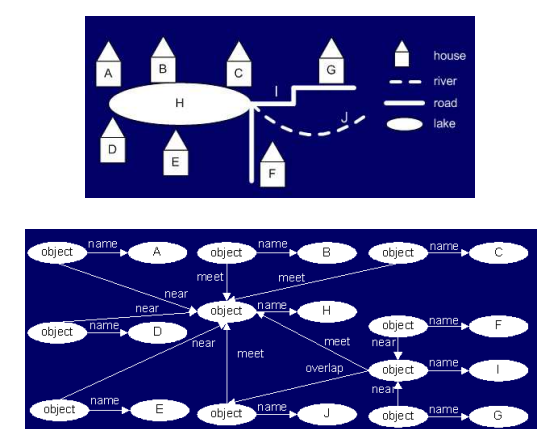
Rappresentazione dei dati

- Vertici: oggetti spaziali ed i loro attributi.
- Spigoli: relazioni spaziali



The diagram shows two 'Spatial object' nodes (circles) connected by three types of relations: 'topological relation', 'direction relation', and 'distance relation'. Each 'Spatial object' node is also connected to two 'value' nodes (circles) via 'attribute' labels.

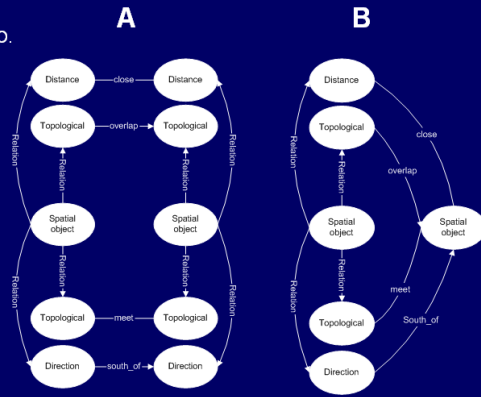
Esempio



The diagram shows a map with various spatial objects (houses, river, road, lake) and their relationships. Below the map is a graph structure representing the spatial data. The graph has nodes for each object (A, B, C, D, E, F, G, H, I, J) and edges representing spatial relations (near, meet, overlap, distance, direction). The legend indicates: house (A, B, C, D, E, F, G), river (H), road (I), lake (J).

Spatial relations:

- Distance: close.
- Topological: overlap.
- Topological: meet.
- Direction: south_of



12.5 – Conclusioni

- Abbiamo tanti database territoriali
- Estrarre il succo dei dati
- Trovare nuove regolarità
- Visualizzazione dei risultati
- Cambio di paradigma
 - passato: scoprire le verità nel mondo reale
 - presente: scoprire le verità nei database