

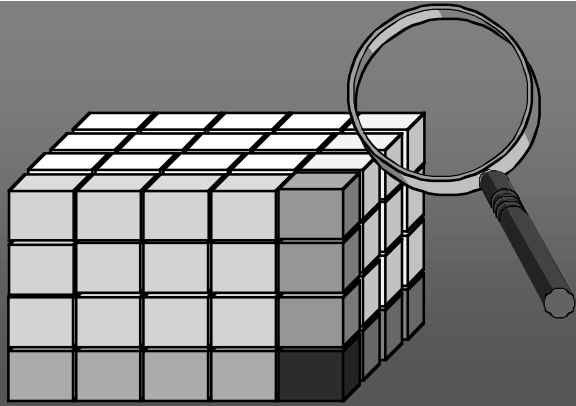
OLAP Mining

OLAP Mining: An Integration of OLAP with Data Mining

Jiawei Han

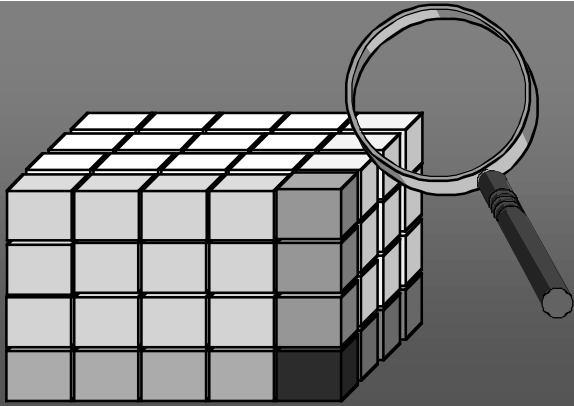
*Simon Fraser University
British Columbia, Canada*

Mandy Whaley 9/15/97



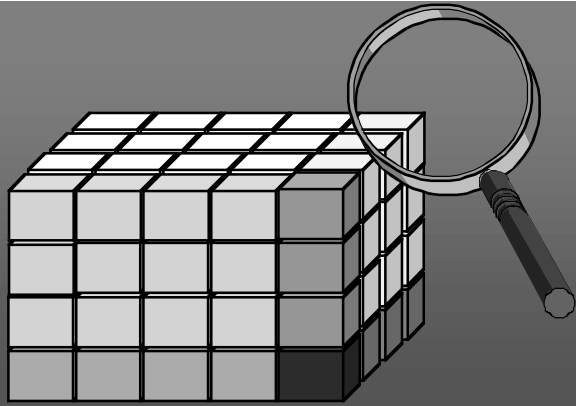
OLAP Mining Definition

“OLAP Mining is a mechanism which integrates on-line analytical processing with data mining so that mining can be performed in different portions of databases or data warehouses and at different levels of abstraction at the user’s finger tips”



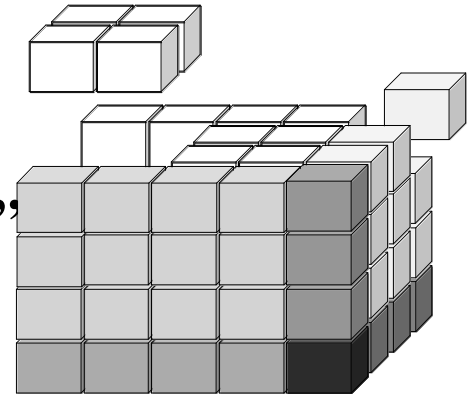
Integration: Why and How?

- **Both Data Mining and OLAP require a clean data warehouse for analysis**
- **OLAP provides a way to view, explore, and summarize data (can be viewed as basic data mining)**
- **Data Mining provides more analysis tools - association, classification, clustering etc.**
- **OLAP/ Data Mining integration allows the user to:
interact with the mining engine based on intermediate results
easily navigate the data and/or the data mining results
select data mining functions dynamically
integrate mining functions -- ex. cluster then associate,
characterized classification**



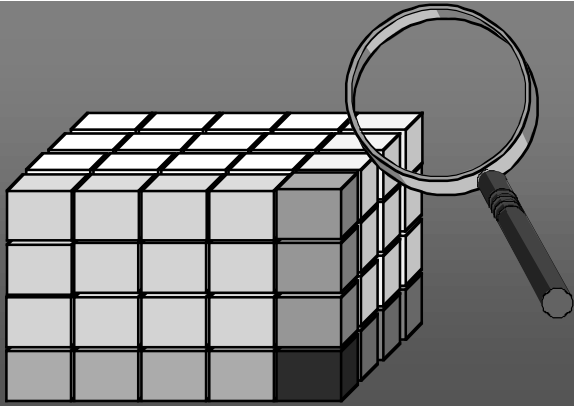
OLAP Operations

**“Data Cube can be viewed as lattice of cuboids.
The bottom most cuboid is the base cube
The top most cuboid only has only one cell.”**



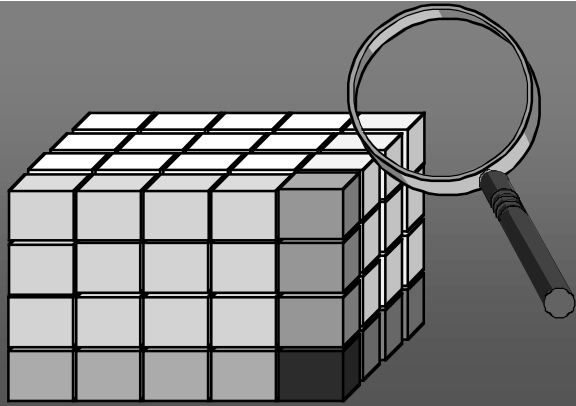
Cubing Operations -- drilling, rolling, slicing, dicing, filtering, pivoting

Each of these operations leads to the generation of new cubes. The new cubes can then be used as a basis for mining.



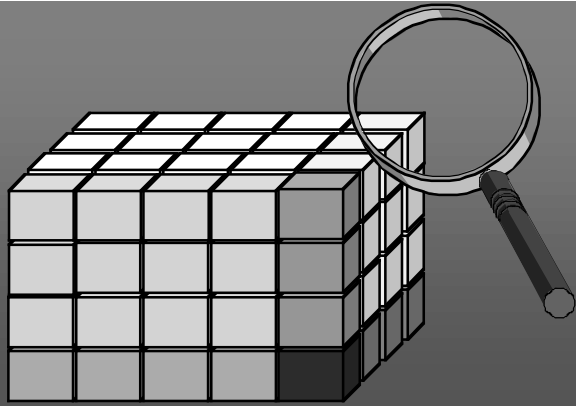
OLAP Functions

- **Cubing then Mining** -- First perform cubing operations to select portion of data or granularity, then start mining process
- **Mining then Cubing** -- Data Mining Results can be analyzed by further cubing e.g. classification then drill down on classes
- **Cubing while Mining** -- Perform Similar Mining Operations at different granularities
- **Backtracking** -- Allows a mining process to backtrack and then explore alternative mining paths.
- **Comparative Mining** -- compare several cluster analysis algorithms by using cubing operations



Examples of OLAP Mining

- **Characterization** - Generalize, summarize, and possibly contrast data characteristics at different abstraction levels, e.g. dry vs.. wet regions.
- **Association** - Discover association rules in the data. e.g a set of symptoms that often occur together with certain diseases.
- **Classification** - Classify data based on values in a classifying attribute. e.g. classify countries based on climate
- **Prediction** - Predict some unknown or missing attribute values based on the other information
- **Clustering** - Cluster data into new classes so that similarity within the cluster is high and similarity between groups is low.

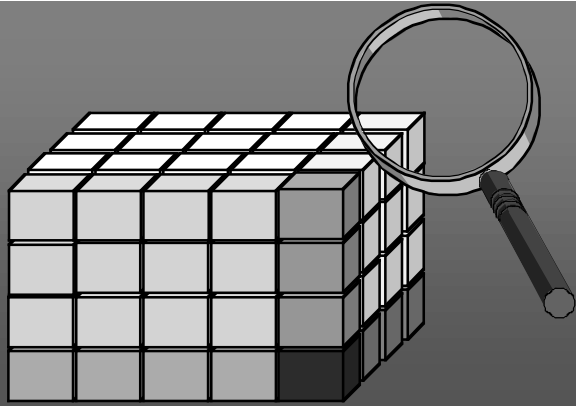


Characterization & Comparison

Characterization --Derives a set of Characteristic Rules which summarize the general rules of a user-specified data set.

Comparison - Derives a set of Discriminant Rules which distinguishes a target class from a contrasting class.

- **Use OLAP features to select data sets.**
- **Can characterize and compare at different levels in the hierarchy.**



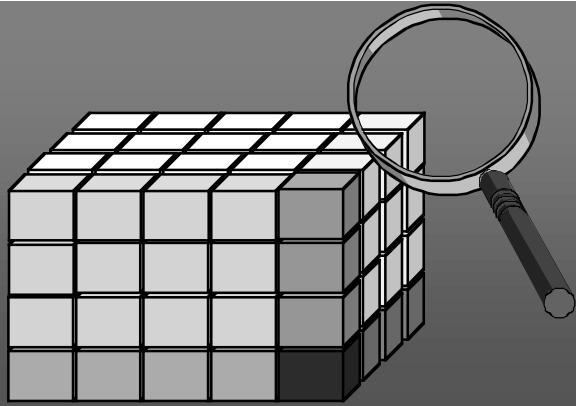
Association

Association - Derives a set of rules about associations between attributes or within attribute sets.

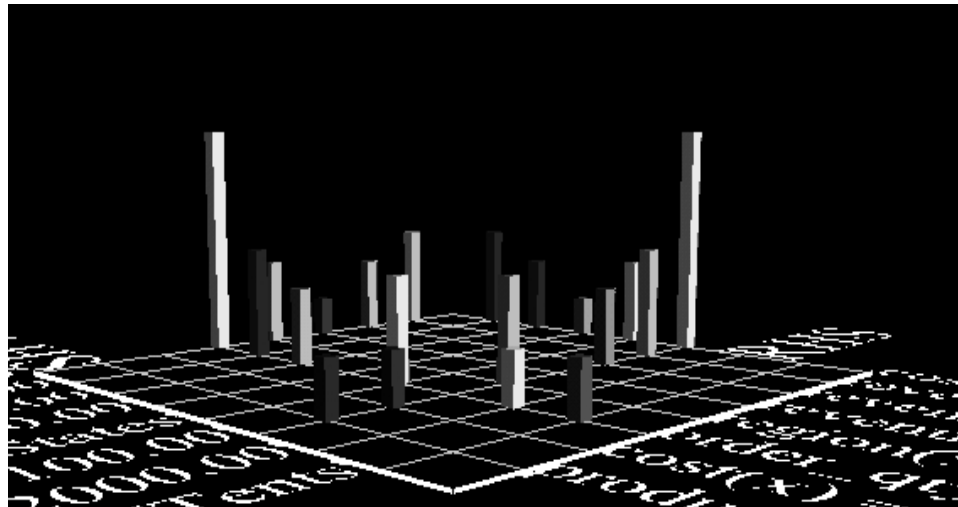
Count Cell - No. of occurrences of the corresponding multi-dimensional data values

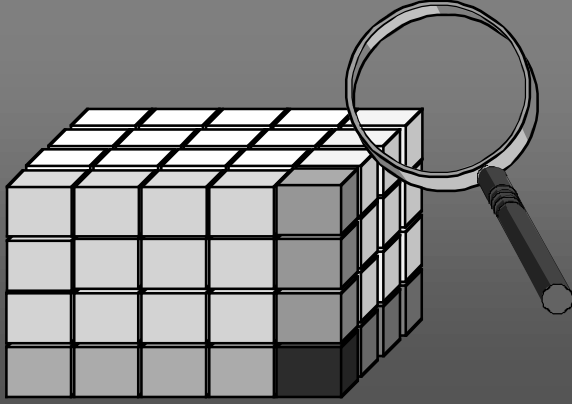
Dimension Count Cell - Sum of counts of the whole dimension

- **Can easily compute support and confidence of association rules using summary cells.**
- **Easy to mine at multiple levels of abstraction.**



Support and Confidence Display





Classification

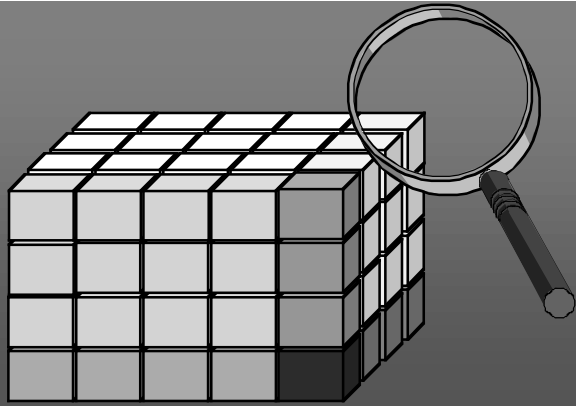
Data classification -- Produces a description or model for each class in a database based on features present in a set of class labeled training data.

- **Minimal Generalization on Training Data**
- **Decision Tree Induction on Generalized Data**

OLAP features help the user select :

an interesting class

the correct level of generalization



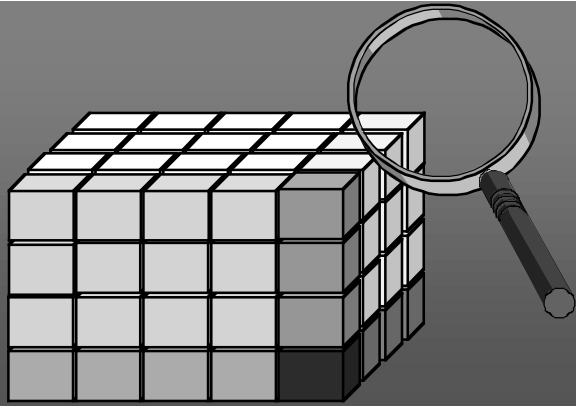
Prediction

Prediction -- Predicts data values or constructs generalized linear models based on the database data.

Dbminer Method

- Minimal Generalization
- Attribute Relevance Analysis
- Generalized Linear Model Construction
- Prediction

OLAP features allow the user to drill-down, slice etc to explore a data cube of predicted results.



Clustering

Clustering -- Partitions a set of data into classes or clusters with high intra-class similarity and low inter-class similarity

Dbminer method:

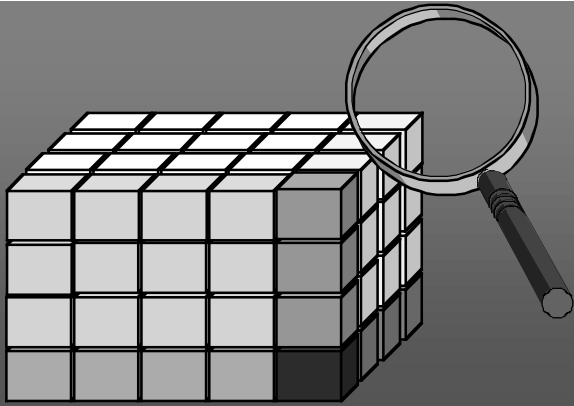
- K-means paradigm

- Added a method for dealing with hierarchies

- Clustering at multiple levels, and select a level by comparing clustering qualities at different levels

- User can direct clustering by assigning weights to attributes

OLAP allows the user to perform cubing operations on clustering results.



Backtracking

“Backtracking is convenient for OLAP mining since a user may like to tentatively dig deep following some mining paths and later try alternatively if there have not been desired interesting patterns found.”

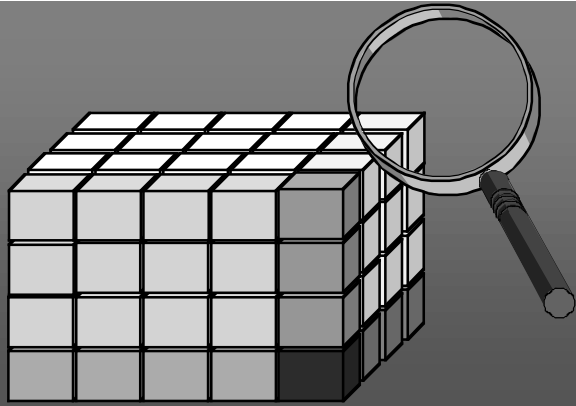
Suggested Method:

Save a status vector in a backtrack stack

The cuboids associated with the vectors should be saved.

Can also compare two mining paths.

When a session is complete, data is cleared.



Conclusions

Integrating OLAP with Data mining gives the user an easy way to navigate and explore the data warehouse and to select portions for data mining. It also allows the user to easily explore the mining results.

Analysis of Paper:

OLAP mining is a reasonable idea. Some of the operations are hard to visualize without using the Dbminer tool. Is the multidimensional cube structure an efficient structure for mining? The backtracking scheme seems to require a bunch of disk space.