CMSC 424 – Database design
Lecture 25
Special databases
Data warehouses
Data mining/Information retrieval

Mihai Pop

Admin

- Course evaluation: http://www.CourseEvalUM.umd.edu
- Review sessions: Thursday & Monday
  - e-mail me topics to cover, questions, problems, etc.

“Special” databases

- Biological data
- Geographic data – GIS
- Movies
- etc.

- New types of queries
- New ways of indexing data
- Storing/retrieval issues (e.g. large sizes, streaming, real-time, etc.)

Examples

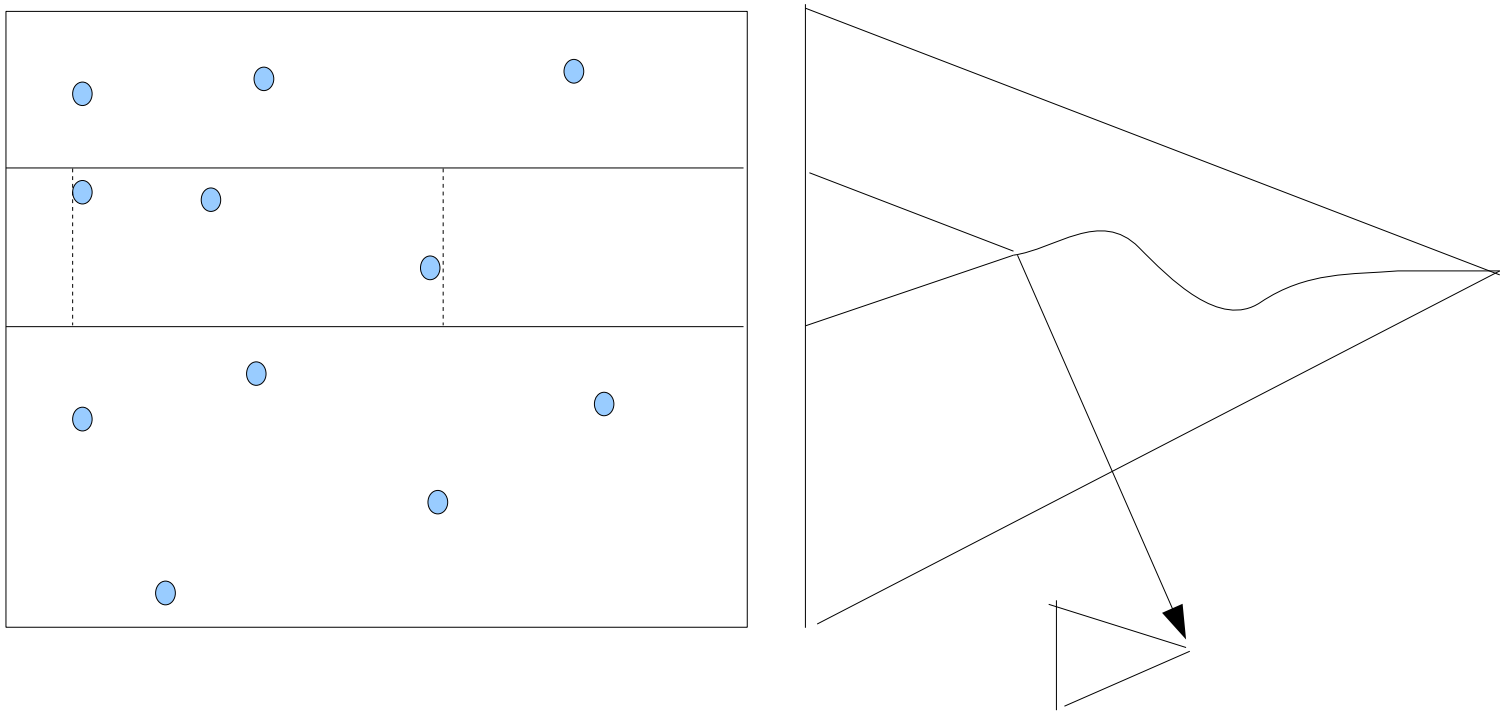
- Biological data
  - refinement of “like” queries: find sequences that are “related”

```
Query: 1   MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVA 60
           M  M++K+L+PTDFSE A  A++  +    ++  EVILLHVIDE  +++     L+ G +
Sbjct: 1   MIFMFRKVLFPTDFSEGAYRAVEVFEKRNKMEVGEVILLHVIDEGTLEE-----LMDGYS 55
```

- Spatial/geographic data (GIS)
  - find all Home Depot stores within 15 miles of Baltimore
  - find a point in Maryland that's farther than 15 miles from the nearest Lowes and is densely populated
  - find all cities within lat/lon square: 39.00 N, 40.00 N, 76.00W, 77.00W.
  - special/spatial index: R-tree

R-tree (chap. 24)

- Binary search tree on Y-coordinate
- Each internal node contains search structure on X-coordinate for all points with Y coordinates in the corresponding subtree



OLAP (chap. 18)

- On-line Analytical Processing
- Why ?
  - Exploratory analysis
    - Interactive
    - Different queries than typical SQL queries
  - Data CUBE
    - A summary structure used for this purpose
      - E.g. *give me total sales by zipcode; now show me total sales by customer employment category*
    - Much much faster than using SQL queries against the raw data
      - The tables are *huge*
- Applications:
  - Sales reporting, Marketing, Forecasting etc etc

Cross Tabulation of *sales* by *item-name* and *color*

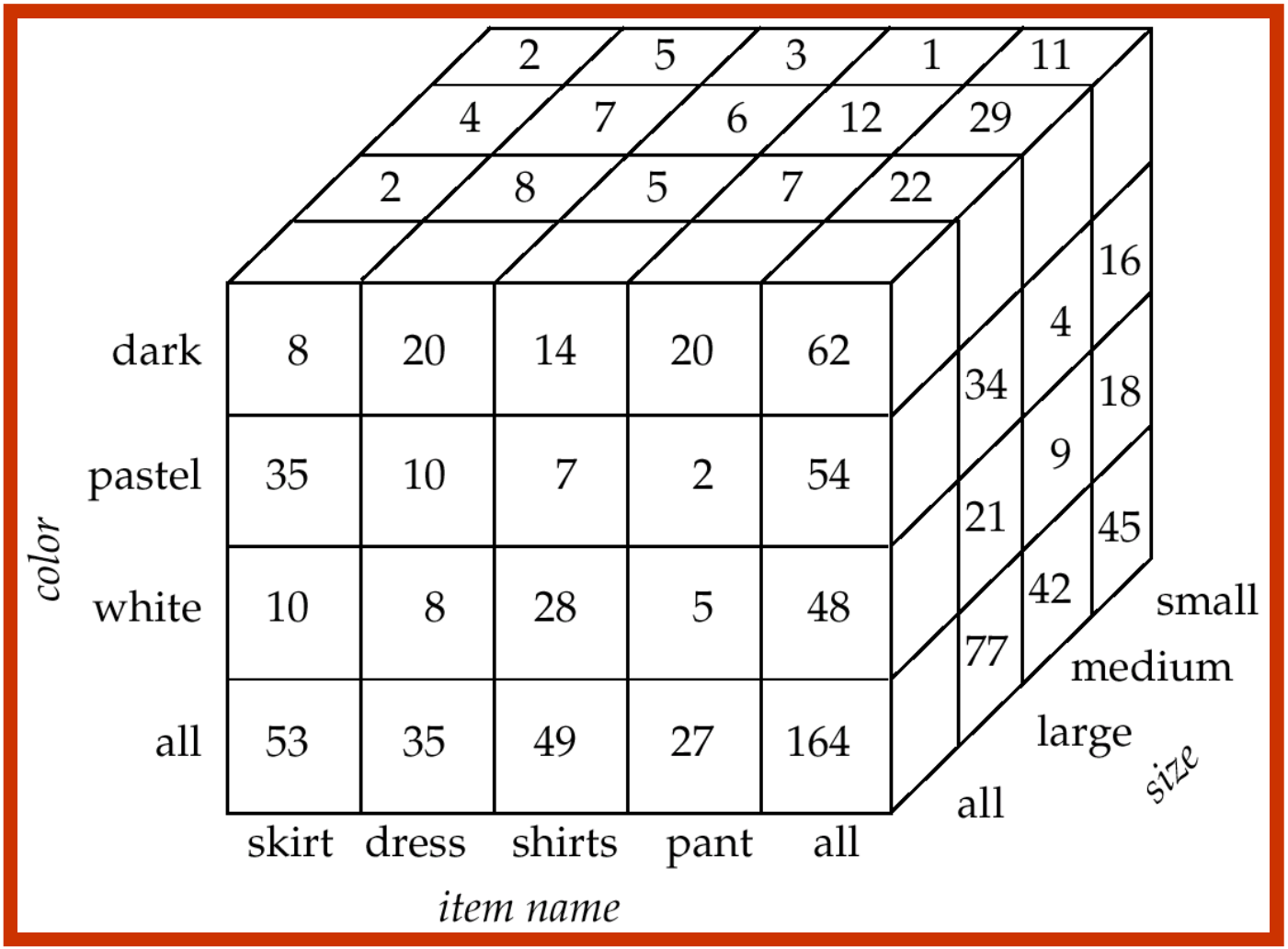
size: **all**

item-name	*color* dark	pastel	white	Total
skirt	8	35	10	53
dress	20	10	5	35
shirt	14	7	28	49
pant	20	2	5	27
Total	62	54	48	164

- The table above is an example of a **cross-tabulation** (**cross-tab**), also referred to as a **pivot-table**.
  - Values for one of the dimension attributes form the row headers
  - Values for another dimension attribute form the column headers
  - Other dimension attributes are listed on top
  - Values in individual cells are (aggregates of) the values of the dimension attributes that specify the cell.

Data Cube

- A **data cube** is a multidimensional generalization of a cross-tab
- Can have *n* dimensions; we show 3 below
- Cross-tabs can be used as views on a data cube



Data federation

- E.g. biological data:
  - VectorBase – organisms that carry human disease (e.g. mosquito)
  - Flybase – fruit flies
  - InsectBase???
- Federation -combining multiple databases into a single virtual database
- Has many issues:
  - schema translation?
  - common vocabulary? (e.g. ontologies, semantic web)
  - privacy/security
  - performance
- Non-biological: SkyServer/SkyQuery (Sloan Digital Sky Survey)

Data warehouses

- Brute-force solution to federation:
  - download all databases
  - convert them to a common schema
  - provide a common interface
- Problems:
  - data storage & duplication
  - hard to keep up to date
  - performance (single point of entry/ failure)
- Examples:
  - GenBank (US biological data repository)
  - Ensembl (EU biological data repository)

Data Mining

- Searching for patterns in data
  - Typically done in data warehouses
- Association Rules:
  - When a customer buys X, she also typically buys Y
  - Use ?
    - Move X and Y together in supermarkets
  - A customer buys a lot of shirts
    - Send him a catalogue of shirts
  - Patterns are not always obvious
    - Classic example: It was observed that men tend to buy *beer* and *diapers* together (may be an urban legend)
- Other types of mining
  - Classification
  - Decision Trees

Information retrieval (chap. 19)

- Extracting **meaning** from **data**
- Examples:
  - Google (document indexing/ranking)
  - Image search
  - Automatic annotation of documents, e.g. extracting information from bio-medical literature

What's next?

- Databases for new types of data (e.g. biological or social networks)
- Streaming databases (Comcast OnDemand)
- Large amounts of data
- Security/Privacy