Review: DBMS Components

Database Management System Components
Research Topics in Databases

- System Oriented…
  - How to implement a DBMS?
  - How to manage the data in the storage?
  - How to construct the index structures for the data?
  - How to implement the different SQL operators?
  - …

- Application Oriented…
  - What kind of database queries can be answered?
  - What kind of index structures can be used to support the database queries?
  - What kind of techniques can be used to improved the database queries?
  - …
A Typical Database Research Project

- Find an **interesting** database research topic: e.g., an useful database query
  - NN query
  - Skyline query
  - Keyword search query
  - …
- Design **“efficient” and “effective”** methods to answer the database query
  - Index structure?
  - Query answering algorithm?
  - Speed-up techniques?
  - Scalability?
- Implementation
  - An executable demo
    - How to maintain the data? How to access data via a database API? How to construct indices?...
  - Conduct experiments on a real dataset
- Publication
  - Write a research paper
    - Publish in major database conferences, e.g., SIGMOD, VLDB, ICDE, EDBT…
  - Apply for a patent
  - Start-up / Commercial software
Next to Discuss

- Database Research
- Data Mining Research
- Web Search and Information Retrieval Research
Query Processing and Advanced Queries

Advanced Queries (1): Spatial Databases and kd-Tree
Spatial Database Applications

- GIS (geographic information system) applications (e.g., maps):
  - Urban planning, route optimization, fire or pollution monitoring, utility networks, etc.
- Other applications:
  - VLSI design, CAD/CAM, model of human brain, etc.
- Traditional applications:
  - Multi-dimensional records
What is a Spatial Database?

- A SDBMS is a DBMS
- It offers spatial data types/data models/query language
  - Support spatial properties/operations
- It supports spatial data types in its implementation
  - Support spatial indexing, algorithms for spatial selection and join
Spatial Data Representation

- Raster model:

- Vector model:
Spatial Data Types

- **Point**: 2 real numbers
- **Line**: sequence of points
- **Region**: area included inside n-points

*Raster and Vector Data*

Raster data are described by a cell grid, one value per cell

- **Raster**
- **Vector**
- **Point**
- **Line**
- **Polygon**
- **Zone of cells**
Spatial Data Relationships

- Topological relationships:
  - adjacent, inside, disjoint, etc
- Direction relationships:
  - Above, below, north_of, etc
- Metric relationships:
  - “distance < 100”
- And operations to express the relationships
Spatial Queries

- **Selection queries:** “Find all objects inside query q”
  - Inside: intersects, north, etc.
- **Nearest Neighbor queries:** “Find the closest object to a query point q”
  - KNN: k-closest objects
- **Skyline queries:** find all skyline points which are not dominated by some others.
Access Methods

- Point Access Methods (PAMs):
  - Index methods for 2 or 3-dimensional points (kd-tree)

- Spatial Access Methods (SAMs):
  - Index methods for 2 or 3-dimensional regions and points (R-tree)
PAM: The problem

- Given a point set and a rectangular query, find the points enclosed in the query
- We allow insertions/deletions online
Tree-based PAMs

- Most of tree-based PAMs are based on kd-tree
- kd-tree is a main memory binary tree for indexing k-dimensional points
- Levels rotate among the dimensions, partitioning the space based on a value for that dimension
- kd-tree is not necessarily balanced
At each level we use a different dimension
kd-Tree Example

Each leaf node can hold up to 2 points.
Spatial Indexing

- Point Access Methods can index only points. What about regions?
  - Use the transformation technique and a PAM
  - New methods: Spatial Access Methods (SAMs)
    - R-tree and variations