COS 425:
Database and Information Management Systems

Professor Andrea LaPaugh

Study how we ***organize*** and ***find*** information

Differences?

- Information
- Knowledge
- Data

What do you want?

A query is a request for data or information satisfying specified constraints

"all students taking Italian"
"information on small villages in Italy"

- Query for information know is there
  Databases
- Query for info will know when see it
  Information Retrieval

– "Surprise me" – *Data Mining*

Data help us?

- Structured data : database
- Semi-structured data: tagged – XML
  HTML?
- Unstructured:
  – Text
  – Other media:
    - Graphics: 2D, 3D
    - Music
    - Video

Structure

But text is structured?
Sentences, subjects, predicates, …

Need *predefined* structure of:

- Types for each basic information object
- *Relationships* between information objects.

That is useful to query/management system

What do you have?

- One big file
- Data you structure
  – database
- Collection of "information objects"
  Examples of "information object":
  - ascii or unicode file
  - HTML file
  - 3D model of a physical object

How do you do it?

- Models of data/information
- Correctness

- In database systems, models of data and correct search well-defined
- In information retrieval, these #1 issues

How do you do it well?

- Organize data storage
- Auxilliary data structures
- Algorithms

Performance issues?

- Large amounts data
  - disk I/O!
- Concurrent use of system
  - Correctness
  - Efficiency
- Distributed across network
  - Where is data?
  - Where should data be?

DB vs IR

Have been looking at ***shared issues***

Now look at ***what distinguishes*** between database (DB) systems and information retrieval (IR) systems

What makes a *database system*?

- Large integrated collection of data
- Uniform access mechanisms
- Model of data organization
  - Levels of abstraction

Database systems ubiquitous
Behind many Web pages

What DB systems provide?

- Uniform interface*
- Uniform models of data*
- Data integrity
- Data security
- Data reliability
- Concurrency
- Efficiency

*like abstract data types but *large:* disk vs memory

Is ***overhead***

Database topics

- Modeling
  - Entity relationship model
    - External "information" view
  - Relational model
    - Conceptional view
    - Foundation of organization and access
  - XML model
    - Databases meet Web

Relational Model

Focus on because dominant DB model
- Formal underpinnings
- SQL most widely used DB language

Historical staying power
Introduced 1970 by Edgar Codd
Flat model
vs older hierarchical and newer XML tree models

Levels of Abstraction

1. Conceptual (e.g. relational) model
2. Data organization
   - indexing
3. Physical model
   - File organization
   - File storage

Determines access and manipulation methods

Database Algorithms

- Data entry
  - Indexing
- Query evaluation
  - requests for data satisfying specified constraints
  - Efficiency
- Achieve concurrency
- Achieve robustness

What makes an *information retrieval system*?

- Large ~~integrated~~ collection of information objects
- Uniform query language
- Model of information object satisfying query

Information retrieval as old a databases
- Gerald Salton SMART project 1960's

Web and large digital collections gave new "life"

Information Retrieval

- User wants information from a collection
- User formulates question as a query
  - usually not exactly capture user need
- System finds objects that "satisfy" query
  - "satisfaction" usually not yes/no but a score
  - Scoring usually not exactly capture user need
- System must present objects to user in "useful form"

Information Retrieval Issues

- Insufficient structure for exact retrieval*
- Best matches versus all matches*
  - What and how present to user?

*not a database system

- algorithms for finding and scoring matches
  - Share indexing techniques with DB

This course and CS fundamentals

Our studies this semester will draw heavily on several fundamental areas of CS:

- Programming methodology
  - semantics of language
  - Correctness
- Algorithms
- Operating systems
  - Concurrency
  - robustness

Course logistics- overview

Web page has all: READ!!
http://www.cs.princeton.edu/courses/archive/fall06/cos425/

- Texts
  - **Required:** *Database Management Systems, Third Edition*, by Ramakrishnan and Gehrke, McGraw-Hill, 2003.
  - reserved books in library
  - online readings
- 2 take-home tests (20% each)
- 6 problem sets (25%)
- Project (30%) – your choosing with approval