COS 597A:

Principles of
Database and Information Systems

Professor Andrea LaPaugh

Database and Information Systems General properties

- Collection of information
- Uniform access mechanisms
- Uniform methods of modifying collection
  - must preserve model

Access mechanisms

A way to get at specific parts of the information.

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

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

- What questions do you want to ask?
- Range of expectations
  - Query for information know is (or isn't) there
  - Query for info will know when see it
  - "Surprise me" – *Data Mining*

Data help us answer?

- Structured data : "database system"
- Semi-structured data: tagged – XML
- Unstructured: "information retrieval sys."
  - 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

Data versus Information?

- Data:
- Information:

Data versus Information?

- Data:
  - fully-defined structure for each item
  - homogeneous database
- Information:
  - "information object"
    - less constraints
    - still has some interpretation
      - ascii or unicode file
      - HTML file
      - 3D model of a physical object
  - heterogeneous collections

How do you answer questions?

- Models of data/information
- Correctness

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

How efficiently to you query/modify ?

- 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?

Have been looking at ***general issues*** of any information system

Now look at details of ***what distinguishes*** between

database (DB) systems and

information retrieval (IR) systems

What makes a *database system*?

- Large integrated collection of data
- Uniform access/modifcation 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
      - conceptional
  - Relational model
    - 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. Logical (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

Our syllubus
Part 1: Models and Queries

- Database models
  - The entity-relationship model
  - The relational model
- XML and the tree model
  - bridging database systems and IR systems
- Information Retrieval

Our syllubus Part 2:
Storing, Retrieving and Maintaining

- Inverted indexes and search
- File Organization
- Indexing Methods
- Relational Query Evaluation
  - Optimization
- Indexes and evaluation for XML
- Transactions

Our syllubus Part 3:
Current Research

- advances in fundamentals and applications
- trend in research:

traditional databases

unstructured information retrieval systems

Graduate Focus

- Emphasize fundamental models and methods
  - expressiveness of languages
  - relationships through constraints
  - effectiveness and efficiency
- De-emphasize how use standard DB systems
  - still opportunity to do so

Graduate Focus

- Explore interaction with "other" research areas
  - research techniques applied to database/info systems
    - example: advanced data structures
    - example: caching in information systems
  - database/info system concepts applied to research
    - example: how integrate heterogeneous data sets in genomics
    - example: how structure data for network monitoring

Course logistics- overview

Web page has all: READ!!

http://www.cs.princeton.edu/courses/archive/fall08/cos597A/

- Texts
  - **Required: *Database System Concepts*** by Silberschatz, Korth, and Sudarshan, 5th Edition, McGraw-Hill, 2006
  - reserved books in library
  - online readings
- 2 take-home tests (15% each)
- 6 problem sets (30%)
- Project (30%) – your choosing with approval
- Class Participation and oral presentation (10%)

* NOTE: will end 5 minutes early for dept. colloquia