Introduction

Zachary G. Ives

University of Pennsylvania

January 13, 2003

CIS 650 – Data Sharing and the Web

Databases vs. Data Management

- “Databases” assume a controlled environment
  - “Closed world” – billing, payroll, etc.
  - DBMS has full control over data
  - Provides guarantees about correctness, etc. (ACID)
  - Based on declarative, logic-based query language
  - … BORING! ☺
- “Data management” uses database-like techniques in a broader context
  - Typically “open world”
  - Data may not be managed internally
  - Fewer guarantees
  - Still aim for declarative queries with logic-based query language

Goals of This Course

- Study data management in a web context:
  - Many providers and consumers of data
  - Uncontrolled environment, local autonomy
- Examine alternative techniques from other fields
  - File synchronization, IR, semantic web, groupware
- Provide a foundation for doing research in this area
- Give each of you experience presenting and analyzing research work

Course Format

- Paper reading, focusing on specific topics
- Combination of lectures by myself, guests, and you
- Course implementation project, presentation, and written report
  - This may be group-based or single-person
- Final examination (take-home)

Your Duties

- Read papers carefully
  - Analyze what they focus on, their major contributions, strengths and weaknesses
  - Compare and contrast with other works, if applicable
  - Post a short 1-page report describing the above to the cis650 newsgroup, due by noon the day the paper is covered in class
- Participate in discussions
- Choose one class day and present the day's readings
  - I'll give you help in preparing, as well as feedback after
- Choose & implement a course project, due by the end of the term
- Take a take-home final exam

My Duties

- Introduce topics and provide necessary background
- Help you prepare for your presentations
- Give you feedback on your presentation
- Evaluate your projects and exams
- Try to answer your questions and facilitate discussions

What We’re Covering

- Most of the papers are in the course reader (to be handed out shortly), but they are subject to change
- Topics include:
  - Data integration – what it does, different systems
  - Query processing – execution, optimization, adaptivity
  - IR-based querying techniques
  - XML processing
  - Answering queries using views – important for schema mediation
  - Versioning, diffs, updates
  - The Semantic Web
  - Groupware

First Assignment

- Today:
  - Choose a paper you’re interested in presenting this semester (I’ll talk more about them next)
  - Tell me now, or send me mail (zives@cis) by Wednesday
  - … Otherwise I’ll assign you a paper arbitrarily!
- By Wednesday:
  - Post to upenn.cis.cis650 a write-up of the TSIMMIS paper (discussed today and very high-level – should be easy!) and the Information Manifold paper by noon

Data Integration and Distributed Data Sharing

- Today and Wednesday:
  - Overview of data integration
  - TSIMMIS: an early semi-structured integration system
  - Information Manifold: defining the mediated schema independently of the sources
- Next Wednesday:
  - Mariposa: a web-scale distributed database using economic model
  - Piazza: a decentralized, peer-to-peer data integration architecture

Query Optimization

- System-R (and a bit on its successor, Starburst)
  - An oldie-but-goody: The 1979 canonical paper on building a cost-based optimizer
- Volcano (and a bit on its predecessor, EXODUS)
  - A rule-based, extensible optimizer that can work on both logical and physical plans

Query Execution

- Graefe’s query execution survey
  - (2 people should read and present this one)
  - Describes relationships between hashing and sorting
  - Describes virtually all of the standard query execution techniques from relational databases

Adaptive Re-Optimization

- Mid-Query Re-Optimization
    - Technique for incrementally optimizing in an RDBMS
- Adaptive Query Execution for Data Integration
    - The Tukwila system and its use of adaptive techniques for data integration
- Eddies
    - A data-flow-based means of adjusting to query costs
- Statistics on Query Expressions
    - Adjusting optimizer estimates based on known results from previous queries

Information Retrieval

- XQuery with keyword search
  - An attempt to bridge IR + databases
- Faloutsos IR survey
  - Survey of techniques for ranking results in IR queries
- WHIRL
  - Approximate joins using IR-style metrics

XML Streams

- (I’ll be at ICDE, so I need student presenters!)
- Tukwila XML query engine
  - First engine to do processing of XML data streams
- Xfilter
  - A publish-subscribe system for XML

Answering Queries Using Views

- Halevy views survey
  - Learn all about inverse rules, the bucket algorithm, MiniCon, and more!
- Schema mediation for P2P
  - How to answer queries in the Piazza system

A Survey of Related Topics

- Change detection for semistructured data
    - “Diff” for unordered XML
- Harmony/Unison
    - (Hopefully a guest lecture)
    - Diffs for files
- Heraclitus
    - A DB programming system based on deltas
- Semantic Web
    - The next big thing?
- Groupware
    - What can we say about the DB problems here?

Tentative Schedule (see syllabus)

Week	Monday	Wednesday
1	1/13: Intro to course; data integration	1/16: Data integration
2	1/20: MLK Holiday	1/22: Mariposa and Piazza
3	1/27: Query optimization	1/29: Query optimization
4	2/3: Query execution	2/5: Query execution
5	2/10: Mid-Query Re-Optimization	2/12: Tukwila
6	2/17: Eddies	2/19: Inter-query adaptivity
7	2/24: XQuery review; Keyword querying	2/26: IR querying, WHIRL
8 (ICDE)	3/3: Tukwila XML	3/5: XFilter
-	3/10: Spring break	
9	3/17: AQUV	3/19: AQUV/P2P Mediation
10	3/24: XML Change Detection	3/26: Harmony/Unison
11	3/31: Heraclitus	4/2: Semantic Web
12	4/7: Semantic Web	4/9: Groupware
13	4/14: Projects	4/16: Projects
14	4/21: Last week of semester -- projects and final exam due	