Database Design and Implementation

CS 645

Course Overview
Resources
Resources
piazza.com/umass/spring2018/cs645

Please sign up and use Piazza for questions of general interest

Please log in to respond to Gradescope consent
Gradescope

Do not create an account!
We will add you
Paper reviews

Do not create an account! We will add you
Database Management Systems
(3rd edition)

http://pages.cs.wisc.edu/~dbbook
Course format

- Mo-Wed, 2:30-3:45pm, Hasbrouck Lab Add 124
- Paper reviews and presentations
- Homework assignments
  - 5 individual assignments
  - group mini-project
- Late Midterm
Disclaimer

- The class is actively designed, so there may be changes to the content, structure, and assignment types.

- You are a crucial part of this development

  - Be vocal about the things you like and the things you don't like

  - Feel free to make suggestions
<table>
<thead>
<tr>
<th>Grading</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework assignments</td>
<td>50%</td>
</tr>
<tr>
<td>Paper reviews, presentations, and class participation</td>
<td>10%</td>
</tr>
<tr>
<td>Midterm</td>
<td>25%</td>
</tr>
<tr>
<td>Mini-project</td>
<td>15%</td>
</tr>
</tbody>
</table>
Course work

- 5 assignments
  - Practical experience
  - Written problem sets
  - Late policy: 3 grace days, 10% penalty per day after that
- 15-18 paper presentations
  - All students need to read the papers
  - Each student presents once in a group of ~5
  - Each student needs to write a review for 3 papers
- Reproducibility group project
Learning goals

- query languages
- relational design
- data modeling
- fundamentals
Learning goals

- fundamentals
- query languages
- relational design
- data modeling
- storage indexing
- processing optimization
- internals
- transactions
Learning goals

- query languages
- relational design
- fundamentals
- data modeling
- storage indexing
- processing optimization
- internals
- transactions

theory
- expressiveness
- complexity
- static analysis

fundamentals
- design
Learning goals

- fundamentals
- query languages
- relational design
- data modeling
- storage indexing
- processing optimization
- internals
- transactions
- expressiveness
- theory
- static analysis
- complexity
- security privacy
- fairness diversity
- advanced
- provenance
- MapReduce
- Spark
Why database research is exciting

- One of the broadest areas
  - Well integrated theory and systems

- A microcosm of CS:
  - Languages, operating systems, data structures, theory, algorithms, distributed systems, statistics
What is a DBMS?

large integrated collection of data

- declarative
- efficient querying
- concurrent users
- reliable storage
- access control
what about file systems?

no efficient access
no query language
no specialized buffering
no recovery from failure
no safe concurrent access
Evolution

- Early DBMSs evolved from file systems
- Many small items, many queries and updates
  - e.g., banking, reservations
- Hierarchical / network model
- Users had to think about how data was stored
the relational model

- E. F. Codd, 1970
- data independence
- declarative language
- mathematical foundation
generality & declarativity

- Programmers and users do not need to know about storage, indexes, sort orders, concurrent users, etc.
- Use logical model, high-level schema
- The DBMS determines how to retrieve the data
levels of abstraction

- View 1
- View 2
- View 3

Conceptual Schema

Physical Schema
Example: university DB

- Conceptual schema:
  - Students(sid:integer, name:string)
  - Courses(cid:integer, name:string, semester:string)
  - Professors(fid:integer, name:string)
designing a schema

- Convert to tables and constraints
- Physical design: disk layout, indices

<table>
<thead>
<tr>
<th>sid</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jill</td>
</tr>
<tr>
<td>2</td>
<td>Bo</td>
</tr>
<tr>
<td>3</td>
<td>Maya</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sid</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>621</td>
</tr>
<tr>
<td>1</td>
<td>645</td>
</tr>
<tr>
<td>3</td>
<td>390</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cid</th>
<th>name</th>
<th>sem.</th>
</tr>
</thead>
<tbody>
<tr>
<td>645</td>
<td>DB</td>
<td>S’18</td>
</tr>
<tr>
<td>621</td>
<td>Soft. Eng.</td>
<td>S’18</td>
</tr>
<tr>
<td>345</td>
<td>DB</td>
<td>F’17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>fid</th>
<th>cid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>621</td>
</tr>
<tr>
<td>2</td>
<td>645</td>
</tr>
<tr>
<td>2</td>
<td>345</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>fid</th>
<th>cid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>621</td>
</tr>
<tr>
<td>2</td>
<td>645</td>
</tr>
<tr>
<td>2</td>
<td>345</td>
</tr>
</tbody>
</table>
find all courses that Mary takes
behind the scenes

SELECT C.name
FROM Students S, takes T, Course C
WHERE S.name = "Mary" AND S.sid = T.sid
AND T.cid = C.cid
DBMSs and DB research

- Huge industry
  - Large data warehouses
  - Distributed databases
  - Integration

- But: not all data is in a DBMS
  - Scientific data
  - Personal data
  - www

- Data management research has expanded
DB research is broad

- core topics (DB internals, processing, optimization, transactions)
- scientific data
- streaming data
- provenance, security, privacy
- cleaning, matching, integration
- distributed data / querying
- usability, visualization
- crowdsourcing
- ...
6:15pm

6:22pm

9:40pm

9:54pm

8:33pm

Tyler Baldwin, Yunyao Li, **Bogdan Alexe**, Ioana Roxana Stanoi: **Automatic Term Ambiguity Detection.** ACL (2) 2013: 804-809

Mauricio A. Hernández, Kirsten Hildrum, Prateek Jain, Rohit Wagle, **Bogdan Alexe**, Rajasekar Krishnamurthy, Ioana Roxana Stanoi, Chitra Venkatramani: **Constructing consumer profiles from social media data.** BigData Conference 2013: 710-716


Bogdan Alexe, Nicolas Heess, Yee Whye Teh, Vittorio Ferrari: **Searching for objects driven by context.** NIPS 2012: 890-898

Bogdan Alexe, Mauricio A. Hernández, Kirsten Hildrum, Rajasekar Krishnamurthy, Georgia Koutrika, Meenakshi Nagarajan, Hagagai Roitman,
MapMerge: Correlating Independent Schema Mappings

Bogdan Alexe
UC Santa Cruz

Mauricio Hernández
IBM Almaden

Lucian Popa
IBM Almaden

Wang-Chiew Tan
IBM Almaden & UC Santa Cruz

Preference-aware Integration of Temporal Data

Bogdan Alexe
IBM Almaden
balexey@us.ibm.com

Mary Roth
IBM Almaden and UCSC
torkroth@us.ibm.com

Wang-Chiew Tan
UCSC
tang@cs.ucsc.edu

Searching for objects driven by context

Bogdan Alexe
BIWI
ETH Zurich

Nicolas Heess
Gatsby Unit
UCL

Yee Whye Teh
Department of Statistics
University of Oxford

Vittorio Ferrari
School of Informatics
University of Edinburgh

Abstract

The dominant visual search paradigm for object class detection is sliding windows. Although simple and effective, it is also wasteful, unnatural and rigidly hardwired. We propose strategies to search for objects which intelligently explore the space of windows by making sequential observations at locations decided based on previous observations. Our strategies adapt to the class being searched and to the content of a particular test image, exploiting context as the statistical
how to debug?

SELECT a3.fname, a3.lname
FROM Actor a0, Casts c0, Casts c1,
     Casts c2, Casts c3, Actor a3
WHERE a0.fname = 'Kevin' AND a0.lname = 'Bacon' AND
     c0.pid = a0.id AND c0.mid = c1.mid AND
     c1.pid = c2.pid AND c2.mid = c3.mid AND
     c3.pid = a3.id AND
    NOT (a3.fname = 'Kevin' and a3.lname = 'Bacon') AND
    NOT EXISTS (SELECT xc1.pid
                 FROM Actor xa0, Casts xc0, Casts xc1
                 WHERE xa0.fname = 'Kevin' AND xa0.lname = 'Bacon' AND
                    xa0.id = xc0.pid AND xc0.mid = xc1.mid AND xc1.pid = a3.id)
GROUP BY a3.id, a3.fname, a3.lname;

alternatives to writing queries?
how do we find a diverse set efficiently?
Questions?

please give us feedback!

CMPSCI 645: Database Design and Implementation

This course covers the design and implementation of traditional relational database systems as well as advanced data management systems. The course will treat fundamental principles of databases such as the relational model, conceptual design, and schema refinement. We will also cover core database implementation issues including storage and indexing, query processing and optimization, and transaction management. Additionally, we will address challenges in modern networked information systems, including data mining, provenance, data stream management, and probabilistic databases.

Course work will include homework assignments, paper reviews and presentations, a (late) midterm, and a mini, collaborative project.

Prerequisites: an undergraduate-level course on databases or operating systems. 3 credits.

Course Time: Mo We 2:30 pm - 3:45 pm, Hasbrouck Lab Add room 124