Course Project Ideas

Yanlei Diao
University of Massachusetts Amherst
New Directions for DB Research

- **Sensor data**: new architecture
- **XML**: new data model
- **Streams**: new execution model
- **Data quality and lineage**: new services
- ...
Querying in Sensor Networks

- Store data locally at sensors and push queries into the sensor network
  - Flash memory energy-efficiency.
  - Limited capabilities of sensor platforms.
Optimize for Flash and Limited RAM

- **Flash Memory Constraints**
  - Data cannot be over-written, only erased
  - Pages can often only be erased in blocks (16-64KB)
  - Unlike magnetic disks, *cannot modify in-place*

- **Challenges:**
  - **Energy**: Organize data on flash to minimize read/write/erase operations
  - **Memory**: Minimize use of memory for flash database.
StonesDB: System Operation

Image Retrieval: Return images taken last month with at least two birds one of which is a bird of type A.

- Identify “best” sensors to forward query.
- Provide hints to reduce search complexity at sensor.
StonesDB: System Operation

Image Retrieval: Return images taken last month with at least two birds one of which is a bird of type A.

Query Engine

Partitioned Access Methods
Research Issues in StonesDB

• Local Database Layer
  – Reduce updates for indexing and aging.
  – New cost models for self-tuning sensor databases.
  – Energy-optimized query processing.
  – Query processing over aged data.

• Distributed Database Layer
  – What summaries are relevant to queries?
  – What remainder queries to send to sensors?
  – What resolution of summaries to cache?
XML: a tagging mechanism to describe content.
XML Data Model (Graph)

Main structure: ordered, labeled tree

References between node: becoming a graph
XQuery: XML Query Language

• A declarative language for querying XML data

• XPath: path expressions
  – Patterns to be matched against an XML graph
  – `/bib/paper[author/lastname='Croft']/title`

• FLOWR expressions
  – Combining matching and restructuring of XML data
  – For $p$ in distinct(document("bib.xml")//publisher)
    Where count($b) > 100
    Order by $p$/name
    Return $p$
Metadata Management using XML

- File Searches
  - all the files generated on Oct 1, 2005
  - all the files whose name is like ‘*simu*.txt’
  - all the files that were generated from the file ‘basic-measures.txt’

- Build an XML store to manage directory trees!
  - XML data model
  - XML Query language
  - XML Indices
XML Document Processing

- Multi-hierarchical XML markup of text documents
  - Multi-hierarchies: part-of-speech, page-line
  - Features in different hierarchies overlap in scope
  - Need a query language & querying mechanism
  - References [Nakov et al., 2005; Iacob & Dekhtyar, 2005]

- Querying and ranking of XML data
  - XML fragments returned as results
  - Fuzzy matches
  - Ranking of matches
  - References [Amer-Yahia et al., 2005; Luo et al., 2003]

- Well-defined problems → identify your contributions!
Data Stream Management

Traditional Database

- Data at rest
- One-shot or periodic queries
- Query-driven execution

Data Stream Processor

- Data in motion, unending
- Continuous, long-running queries
- Data-driven execution
In-Network XML Processing

• XML is becoming the wire format for data
• In-network XML processing
  – Authentication
  – Authorization
  – Routing
  – Transformation
  – Pattern matching

• XPath widely used for in-network XML processing
  • Applied directly to streaming XML data
  • Line-speed performance
Research Issues

- **Gigabit rate XPath processing**
  - Take one look, process XPath, buffer data for future use if necessary
  - Processing needs to be gigabit rate
  - Memory usage needs to be minimized

- **Time/space complexity of XPath stream processing**
  - Theoretical analysis for common features of XPath

- **Minimizing memory usage of YFilter technology**
  - YFilter: state-of-the-art for multi-XPath processing
RFID Technology

- RFID technology

reader_id, tag_id, timestamp

Yanlei Diao, University of Massachusetts Amherst
RFID Stream Processing

<pm1>
<tag>01.01298.6EF.0A</tag>
<time>00129038</time>
<location>shelf 2</location>
</pm1>

+ RFID tag

RFID reader

<pm1>
<tag>01.01298.6EF.0A</tag>
<time>02183947</time>
<location>exit1</location>
</pm1>

Yanlei Diao, University of Massachusetts Amherst 2/6/2007
Example Queries

• **Shoplifting**: an item was taken out of store without being checked out.

• **Out of stocks**: the number of items of product X on shelf ≤ 3.

• **Misplacement**: an item was moved from Shelf A to Shelf B without being purchased or put back.

• …
RFID Processing: Global Tracking

Yanlei Diao, University of Massachusetts Amherst 2/6/2007
Example Queries

- **Counterfeit drugs**: a bottle is accepted at the retailer if it came from a legal manufacturer and followed all necessary steps in the distribution network.

- **Expired/spoiled drugs**: a bottle is accepted at the retailer if it went through the distribution network in less than 3 months and was never exposed to temperature > 96 F.

- **Missing pallet, expected case, illegally cloned tags…**
Challenges in RFID Management

• Data-Information Mismatch
  – RFID raw data: (tag id, reader id, timestamp)
  – Meaningful information: shoplifting, misplaced inventory, out-of-stocks; expired drugs, spoiled drugs…

• Incomplete, inaccurate data
  – Readers miss tags
  – Readers can pick up tags from overlapping areas

• High-volume data
  – Readers read constantly, from all tags in range, without line-of-sight
  – Can create up to millions of terabytes of data in a single day

• Low-latency processing
  – Up-to-the-second information, time-critical actions
Research Issues

• Real-time event stream processing
  – Handling duplicate readings/results
  – Data cleaning
  – Data compression

• Handling incomplete readings
  – Inferences in event databases
  – Inferences over event streams

• Distributed processing
  – Real time anomaly detection
  – Distributed inferences
Adaptive Sensing of Atmosphere

- Environmental monitoring: real-time processing of huge-volume meteorological data

- **Challenges**
  - Large volume but limited bandwidth
  - Real-time processing
  - Uncertain data
  - Data archiving and querying the history
Managing Uncertain Data

- **Sources of data uncertainty**
  1) Sensing noise and partial scanning
  2) Data compression
  3) Lossy wireless links
  4) Incomplete merging

- **Managing uncertain data**
  - Model sources of data uncertainty
  - Develop uncertainty calculus to combine the effects of these sources
  - Augment results with confidence values
Managing Uncertain Data

- **Sources of data uncertainty**
  1. Sensing noise and partial scanning
  2. Data compression
  3. Lossy wireless links
  4. Incomplete merging

- **Self diagnosis and tuning**
  - Compare prediction at $t$ with observation at $t+1$ (no ground truth?!) 
  - System diagnosis when confidence value is low
  - Automatically tune the system
Questions