Modern Data Management

Tingjian Ge
ge@cs.uml.edu
The problems we solve

Holter Monitors

Sensors attached to chest of a heart-disease patient

Continuously monitor the ECG signals

Sent to a server through wireless communication
Holter Monitors

R-R interval too short: match $p = 3003$ with a small window size $w_1$

R-R interval too long: match $p = 3003$ with a large window size $w_2$
The problems we solve: Highway Monitoring

Scenario 1: finding close-by tailgaters.
Sensors can measure gaps between two cars.

T – truck tailgater, C – car tailgater, N – normal vehicles. We get a real-time sequence stream.

“All alert me when there are at least 5 truck tailgaters within a window of 20 vehicles.”

p = TTTTT, window size = 20.

Scenario 2: detecting a sudden traffic jam.
Sensors can estimate speed of vehicles passing by.
Discretized into 0 (<10 mph), 1 (10 to 30 mph), 2 (30 to 50 mph), 3 (> 50 mph)

p = 3210 within a certain window size – indicating a sudden slow down.
The problems we solve: Smartphones and RFID

Location estimates using WiFi, GPS, …

Shopping mall: customer location + map = noisy stream of each customer’s visit sequence

A business owner’s query:
\[ p = <\text{restaurant}, \text{dessert shop}, \text{movie theater}> \]

President Barack Obama joined foursquare in 2011.

RFID on shopping cart: shopper’s shopping path
-- business intelligence

RFID on shelves, etc.: detecting shoplifting:
\[ p = <\text{shelf}, \text{not register}, \text{exit}> \]
Examples of real systems

Scientific Data and Sensor Networks

Social Network Service

Data Integration

Data Extraction
Data Stream Management System (DSMS)

- Users register queries
- High rate data flowing in
- Query results, notifications, warnings
- Scratch Store
- Lookup Tables
- Archive
- Online Performance Metrics
Q1: SELECT name FROM employees WHERE salary > 80000

Q2: SELECT AVG(salary) FROM employees WHERE salary > 60000
Data Management in Cloud Computing

Parallel and distributed computing

Processing large scale data:
• E.g., scientific data
• Parallel databases
• MapReduce framework used by Google, Yahoo,…
Project *Rural* (Querying Rich uncertain data in real time)
Project **MUSE** (Managing Uncertain Scientific Experimental Data)

[Diagram showing the structure of MUSE, including high rate sequence data feeding into CPM (continuous pattern monitor), and relational and sequence data feeding into POPE (parser, optimizer, plan generator, executor). Outcomes include one-time query, monitoring query, and results, with PAM (probability & accuracy manager) and Data Dictionary as components.]
Thank you. Questions?