More Model-Driven Processing in Sensor Networks

Jun Yang
CPS 296.1, Spring 2007
Sensor Data Processing
With contents from A. Deshpande

Announcements (Feb. 8)

- Course project milestone 1: March 1
- Reading for next Tuesday (review): directed diffusion

More model-driven processing

- Amol Deshpande, Carlos Guestrin, Samuel Madden, Joseph M. Hellerstein, and Wei Hong. “Model-Driven Data Acquisition in Sensor Networks.” International Conference on Data Engineering, 2006
- Yannis Kotidis. “Snapshot Queries: Towards Data-Centric Sensor Networks.” International Conference on Data Engineering, 2005
Ken

- Forget about fancy queries—many sensor network users still want to collect all data
- As in BBQ: exploit correlations in data
  - What's different from BBQ?
    - Use model to optimize transmission, but not to avoid acquisition
    - Don’t transmit if reading predicted by model

Models for prediction

- I know past samples from this node: \(X_i, X_{i-2}, X_{i-1}, X_{i+k}, X_{i+m}, X_{i+n}, \ldots\)
- My guess for the new sensor samples is… \(X_{i+1}\)

\[\text{model (input \ldots)} \rightarrow \text{predicted value}\]

A simple example

(Transition) model: value predicted at \(t+1\) \(\leftarrow\) value predicted at \(t\)

- Source reporting rule:
  - report if \(|\text{actual} - \text{predicted}| \geq \varepsilon\)
- Model update rule:
  - use actual if report received
- Guarantee:
  - \(|\text{actual} - \text{predicted}| < \varepsilon\)

Another possibility: A linear model

\[X_{t+1} = a X_t + b\]
From temporal to spatiotemporal

From temporal to spatiotemporal

1 node
no spatial correlations

entire network
full spatial correlations

Ken explores the spectrum between these two

What's the trade-off?

Some details
In each time step...
- Apply transition model to get \( p(X_{t+1}) \), conditioned on all observations transmitted to sink
  - Caution: not all observations!
- Source:
  - Compute \( X_{t+1} \) using \( p(X_{t+1}) \)
  - \( \forall i: |x_i^{t+1} - x_i^t| < \varepsilon \), stop
  - Otherwise, send to sink the smallest subset of values such that conditioning on it makes predictions accurate
- Sink:
  - Condition \( p \) on values received from source (if any)
  - Compute \( X_{t+1} \) using \( p(X_{t+1}) \)

Example: disjoint cliques
- Nodes form cliques with roots
- In each time step
  - Nodes in a clique transmit to clique root
  - Clique root runs model (mirrored at base station) and decides which readings (if any) to transmit to base station
Optimizing disjoint cliques

- What to consider
  - 
  - 

- How to compute data reduction factor
  - Use the model (poor model gives suboptimal cliques)
  - Done compile-time, at base station
    - NP-hard; use exhaustive or heuristic search

Example: average model

When would this model work well?

Experiment results

- Ignores model maintenance (why?)
Discussion

- Main concern of BBQ addressed
  - Suppression driven by in-network models
  - No longer trust model completely
  - Can handle outliers
  - Though detecting node failures again depends on good models

Snapshot queries

- Target one-time queries instead of continuous ones

Ideas

- Use a subset of nodes (representatives) to represent their neighbors to within ±T
  - Queries contact fewer nodes
- Rep N_i has a model for each N_j it represents
  - Requires communication between N_i and N_j to maintain

Model maintenance

- Rep N_j “caches” recent readings from N_j
- Model obtained by fitting a line through all available (x_i’, x_j’) pairs: x_j’ = a_i,j x_i’ + b_i,j
- Cache replacement policy considers benefit of caching a pair and cost of evicting a pair, in terms of prediction error
Selecting and maintaining reps

- Rep discovery is fully distributed
  - Nodes broadcast values & signal they're looking for reps
  - Nodes with values within $\pm T$ of $x_j$ add $N_j$ as a candidate
  - Nodes broadcast their lists of candidates
  - Node $N_j$ picks the node with the longest list containing $N_j$ to be its rep, and let that node know
  - Some refinement logic
- A node $N_j$ represented by $N_i$ periodically sends a heartbeat message to $N_i$ asking for its estimate
  - If estimate out of bound or no reply, select a new rep

Comparison with BBQ/Ken

- How are the models used?
- Where are the models?
- Are models replicated?
- How complex are the models?
- How are the models built?
- Who answers queries?