Accuracy-aware Data Modeling in Sensor Networks

Ryo Sugihara

Andrew A. Chien

Computer Science and Engineering Dept, Center for Networked Systems, University of California San Diego

{ryo, achien}@ucsd.edu

Motivation
- Accuracy of data in sensor networks
  - Should be under control of applications
- Existing approaches: best-effort
  - No guarantee on accuracy

Problem: Pick an efficient data representation that meets the specified accuracy requirement

Accuracy-aware data modeling

The Problem
- Assumptions
  - Sensor nodes are separated geographically into disjoint groups
  - Remote host communicates with all leaders and acts as a gateway to remote host to send data
  - Leader communicates with all sensors in the group

- Application is interested in an environmental attribute (e.g. temperature) over the entire space
- Application has an accuracy requirement
  - Expressed as an “Error bound”
- Optimization
  - Objective: Compact representation of data
    - (Coordinator)
  - Constraint: Must satisfy accuracy requirement
    - (Represented within the error bound $\epsilon$)

Key Idea: Choose cutoff to meet accuracy requirement

Approach
- Hybrid: Regression Plane + Explicit Points

Data summarization and filtering

For each group: do iteratively (@ Leader)
1. Calculate regression plane for data $f_i(x,y) = ax + by + c$
   $\begin{align*}
   \sum_{i=1}^{n} & \left( z_i - (ax_i + by_i + c) \right)^2 \\
   \Rightarrow & \quad \min \\
   \Rightarrow & \quad a \text{ and } b \text{ and } c 
   \end{align*}
   $
   \begin{align*}
   \Rightarrow & \quad a \text{ and } b \text{ and } c 
   \end{align*}$

2. Filter out a data
   - Data with the maximum deviation $\delta = \max_{i} (|f_i(x,y)| - 1)$
   - $d_i \leq \delta$, break the loop
   - Otherwise eliminate sensor #m and repeat

Combining regression planes
- Iteratively combine regression planes (@ Coordinator)

Collecting explicit data
- Coordinator requests explicit data for each group leader
- Leader sends raw data back to coordinator

Simulation Experiment
- Simulation data:
  - Environmental data
    - Brightness temperature
  - 155 sensors, 16 groups

   - Resulting data representation:
     - $\epsilon$ (without summarization) produces more complex representation
     - $\epsilon_1.5$ (with summarization) produces efficient representation

Results
- Flexible tuning of accuracy
- Fairly good performance

Conclusion & Future Work
- Accuracy-efficiency trade-off
  - Assuring accuracy is first-priority
    - Previous approaches focus on efficiency only
  - Our approach provides a new spectrum of flexibility
- Future work
  - Scalability: Fully-distributed algorithm (Ongoing)

Supported in part by the National Science Foundation under awards NSF Cooperative Agreement ANI-0225642 (OptIPuter), NSF CER-0332446 (VGG-AD), NSF ACI 0095790, and NSF Research Infrastructure Grant EIA-0105622. Support from the UCSD Center for Networked Systems, BepKidsheets, and Fujitsu is also gratefully acknowledged. Support from IBM Japan is also gratefully acknowledged.