


Time Series Forecasting in Answering Queries and Detecting Data Similarities in Wireless Sensor Networks

Daniela Tulone
MIT CSAIL
tulone@csail.mit.edu
(joint work with Sam Madden)

10/2/2006

Introduction

- WSN offer a potential to collect large amount of data from remote locations
 - large variety of applications: environmental and industrial monitoring, agriculture...
- Data is often collected at sink, and analyzed
- Tools to facilitate data collection (e.g., Cougar, TinyDB, Directed Diffusion, etc...)

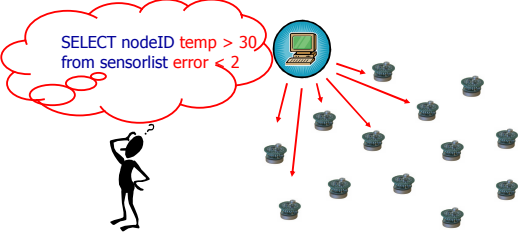


10/2/2006

The problem

Problem: **efficiently** answer queries at sink

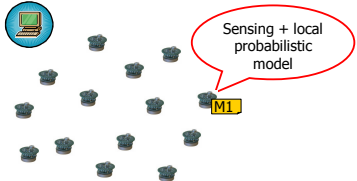
SELECT nodeID temp > 30
from sensorlist error < 2



10/2/2006

Our approach

- **Goal:** conserve energy by reducing transmission
- **Approach:** probabilistic
 - local/global models
 - linear time series → compact representation of data



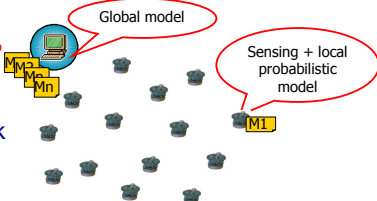
10/2/2006

Our approach

- **Goal:** reduce communication
- **Approach:** probabilistic
 - local/global models

SELECT nodeID temp > 30
from sensorlist conf > 95%
error < 2

Global model at sink
no communications



10/2/2006

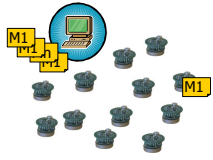
Contributions

- SAF: framework for approximate **query answering** and detecting **data similarities**
 - Class of **lightweight adaptable** time series models
 - Centralized query framework based on **local models**
 - Detect outliers, data variations and inconsistencies
 - Detect **node similarity** and group nodes into **clusters** at **no additional cost**
 - **Novel definition** of data similarity based on **prediction values** → very efficient clustering algorithm, **optimal** in the number of clusters
 - Features: **energy-efficient**, **robust**, suitable for **mobile networks**, provide useful information, **adaptable**, provably **correct**, **general**
 - Analytical and experimental evaluation

10/2/2006

Outline of the talk

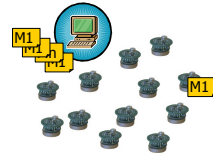
- Local probabilistic model
 - Simple time series model, motivations
 - learning and monitoring algorithms
- Centralized query framework
 - query answers
 - node similarity
 - simulation results



10/2/2006

Outline of the talk

- Local probabilistic model
 - Simple time series model, motivations
 - learning and monitoring algorithms
- Centralized query framework
 - query answers
 - Node similarity
 - Simulation results



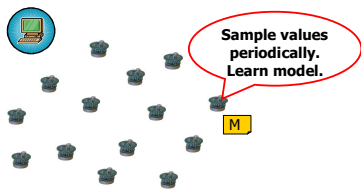
10/2/2006

SAF overview

Phase 1

Steps Phase 1:

- Sample values every T time units
- Learn model



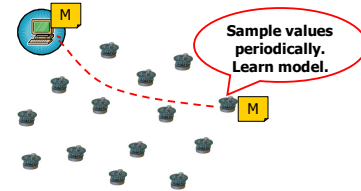
10/2/2006

SAF overview

Phase 1

Steps Phase 1:

- Sample values every T time units
- Learn model
- Transmit model to sink



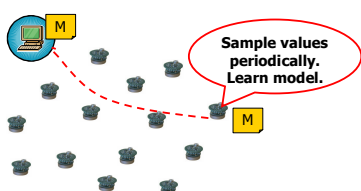
10/2/2006

SAF overview

Phase 1

Sink stores local models and answers query

- Model has to fit data distribution
- Copy at the sink in synch with local copy



10/2/2006

SAF overview

→ Data distribution can change over time!

10/2/2006

SAF overview

➔ Data distribution can change over time

Steps Phase 2: **Phase 2**

1. Sample values every T time units
2. Monitor error

10/2/2006

SAF overview

➔ Data distribution can change over time

Steps Phase 2: **Phase 2**

1. Sample values every T time units
2. Monitor error
3. If needed re-learn model

10/2/2006

SAF overview

➔ Data distribution can change over time

Steps Phase 2: **Phase 2**

1. Sample values every T time units
2. Monitor error
3. If needed re-learn model
4. Transmit new model to sink

10/2/2006

Time series forecasting

- Well-known statistical technique
 - > finance, network traffic, dynamic resource allocation
- Time series = set of temporal observations
 - > predict current value based on previous history

- Linear time series models → ARMA
 - Expensive to built at sensor nodes!
 - long training phase
 - high memory and computational cost

10/2/2006

Our model

- Model challenges in WSN:
 - > lightweight model with short learning phase
 - > accurate, robust to data noise

10/2/2006

Our model

- Model challenges in WSN:
 - > lightweight model with short learning phase
 - > accurate, robust to data noise
- Subclass of ARMA models
 - > AR(q) model → prediction based on last q values
$$X(t) = a X(t-1) + b X(t-2) + c X(t-3) + d N(0,1)$$

10/2/2006

Modeling physical phenomena

Observation: phenomena usually change **slowly!**

Idea:

- Don't try to model a complex phenomenon $F!$
- Compute an **adaptable lightweight** model M capable of accurately predicting F during some time window W
- Model F is: $M = \{ M_1, M_2, \dots, M_i, \dots \}$

$$F(t) = m(t) + X(t)$$

- Ex. $P(t) = m(t) + a X(t-1) + b X(t-2) + c X(t-3)$

10/2/2006

Metrics

- Model accuracy \rightarrow **uncertainty** and **error probability**

Lemma Prediction error $|P(t)-v|$ is smaller than $e=a*d$, with error probability smaller than $1/a^2$.

- Cost of learning/adapting
- Communication cost: send **new model coefficients** + ~~periodic readings~~

10/2/2006

Model monitoring

- Periodically read value v ,
- Compute $err=|P(t) - v|$, $e = a * d$

10/2/2006

Outline SAF presentation

- Local probabilistic model
 - AR model, motivations
 - learning and monitoring algorithms
- Centralized query framework**
 - query answers
 - Node similarity
 - Simulation results

10/2/2006

Answering user queries

SELECT nodeID temp > 30 from sensorlist error < 1

10/2/2006

Answering user queries

SELECT nodeID temp > 30 from sensorlist error < 1

Remarks:
1. Avoid periodic readings:
bound sensor readings at sink using stationarity of X(t)

10/2/2006

Answering user queries

SELECT nodeID temp > 30 from sensorlist error < 1

Remarks:
Avoid periodic readings:
bound sensor readings at sink using stationarity of X(t)
Tunable rate
Dynamic time series model

10/2/2006

Answering queries

SELECT nodeID temp > 30 from sensorlist error < 1 conf > 95%

Tunable according to application requirements

```

Query(Q)
for each node i do
  if Q.err > err_i & trend(t) satisfies Q.cond
    add i into query sensorlist L
  else
    add N into sendlist S
  Contact nodes in S and unstable list
  
```

Query(Q) is provably correct.

10/2/2006

Data similarity

- Detecting data similarities is relevant for a number of sensor applications:
 - Redundancy detection → low duty cycle, or replacement of sensors
 - Intrusion and anomaly detection
 - Scientific studies, detect spatial-temporal correlations
 - Volcanic applications, etc...
- Clustering under **dynamic conditions** is energy-consuming!
 - Computing and maintaining clusters involves **coordination among nodes**
 - PAQ System [EWSN 06]
- Idea: detect similarity not based on raw data but on **data models**

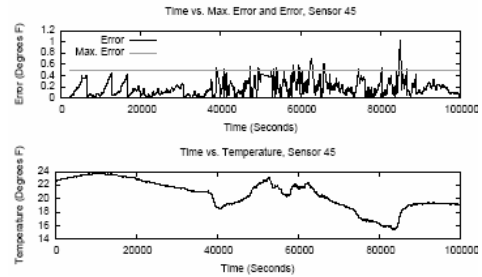
10/2/2006

Data similarity

- Detecting data similarity based on prediction values:
node i and j are **a-similar** if $|P_i(t) - P_j(t)| < \alpha$
- Bound prediction value $P(t) \rightarrow [l, L]$
- Benefits:
 - transform complex problem into 1-dim problem
 - clustering algorithm $O(n \log n)$, **optimal** number of clusters
 - larger clusters than geographical clusters
 - nodes are not aware membership
 - variations in the cluster formation do not involve **additional communication cost**
 - suitable for **mobile networks!**

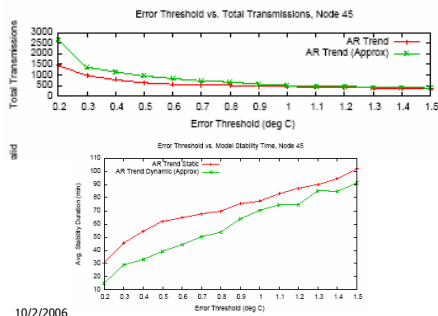
10/2/2006

Model stability



10/2/2006

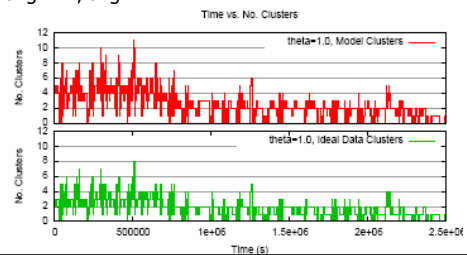
User-error vs. transmissions



10/2/2006

Clustering

- Number of clusters over the system lifetime: 50 sensors, similarity within 1 degree
- Comparison between ideal clusters and our approach: avg. 2.4, avg. 4.2



SAF features

- Consumes little **energy**
 - Reduced communication + low duty cycle
 - Short learning phase
- Suitable for **mobile networks**
 - More robust to communication failures
 - No additional overhead
- Ability of deriving **strong data properties**
 - Outlier detection, data inconsistencies, data similarities
- **Adaptable** strategies
 - **Tunable data rate**, areas of geographical interest

10/2/2006

Comparisons

	SAF	BBQ	Ken	Kalmar
Adaptability	Yes	No	Yes	Yes
Robust to comm. fail.	Yes	No	No	Yes
Outlier	Yes	No	Yes	No
Learning	1/2 - 1 h	7-15 days	few days	—
Transmissions over week	60	> 20160	>2880	~60
Data similarities	Yes	correlations	correlations	No

10/2/2006



Applications

- General framework
 - Approximate data stream evolving over the time
 - Highly dynamic and limited settings
- Wireless sensor networks
 - Time synchronization
 - Data integrity
 - Data calibration
- Sensor networks and the Web
- Load balancing, intrusion/anomaly detection...

10/2/2006



Conclusions

- Proposed SAF framework for approximate query answer and outlier and similarity detection
 - **Lightweight** and **adaptable** time series models
 - Showed suitability simple model + monitor
- Main features of SAF:
 - Consume **little energy**
 - Provide information regarding outliers, variations in the data distribution, periods of data inconsistency
 - Provably correct
 - Adaptable: provides user-controllable error
 - Suitable for mobile networks
- Further applications include data calibration, object tracking, security...

10/2/2006