2009 15th International Conference on Parallel and Distributed Systems

Energy-Efficient Sensor Data Acquisition based on Periodic Patterns

Guan-Rong Lin[1], Yao-Chung Fan[2], En Tzu Wang[2], Tao Zou[3], Arbee L.P. Chen[4]*
[1]Industrial Technology Research Institute, Taiwan, R. O. C.
[2]National Tsing Hua University, Taiwan, R. O. C.
[3]Fudan University, P. R. C.
[4]National Chengchi University, Taiwan, R. O. C.
alpchen@cs.nccu.edu.tw[4]

Abstract—**Wireless sensor networks have received considerable attention in recent years and played an important role in data collection applications. Sensor nodes usually have limited supply of energy. Therefore, a major consideration for developing sensor network applications is to conserve the energy for sensor nodes. In this paper, we propose a novel energy-efficient data acquisition algorithm based on the periodic patterns derived from past sensor readings. Our key observation is that sensor readings often exhibit periodic patterns, e.g., the daily cycle of temperature readings, and the patterns provide opportunities for reducing energy consumption for sensor data acquisition. We exploit the patterns and use the patterns to build a statistic model for predicting sensor readings. In our approach, sensor data acquisition is needed only when acquired readings are unpredictable. Therefore the energy for sensor data acquisition and the associated radio communications can be conserved. The experiments performed with real data validate the effectiveness and efficiency of our approach.**

Keywords: Sensor Networks, Data, Acquisitions, Query Processing.

I. Introduction

Wireless sensor networks have received considerable attention in recent years and played an important role in data collection applications. A wireless sensor network typically consists of a large number of sensor nodes equipped with the abilities of sensing, computing, and communicating. Wireless sensor networks provide new means for collecting data. One promising application of sensor networks is the scientific data collection, in which sensor nodes are deployed in the field where data are difficult or expensive to collect. In such data collection applications, sensor nodes periodically take sensor readings to produce a dataset for offline scientific analysis.

One of the features for wireless sensor networks is resource limitations. Sensor nodes typically are limited in computing power, network bandwidth, storage capability, and energy supply. The limited computing and storage capability restrict the data processing algorithm that can be operated on the sensor nodes. In addition, the sensor nodes become useless once the batteries are depleted. Reinstalling the batteries for hundreds of sensor nodes is labor-intensive and impractical. Therefore, resource conservation becomes a major consideration when devising sensor applications.

*To whom all correspondence should be sent.

In continuous data collections, sensor data often exhibit patterns which provide opportunities for reducing the energy consumption for the uses of sensor nodes. There has been a great deal of interests in recent years in developing approaches exploiting the patterns to reduce energy consumption. In [9], an energy-efficient querying framework is built based on the observation that sensor readings often change infrequently. The basic idea behind the approach is that a sensor node sends its reading only when the reading significantly differs from the previous reading. The energy for communications between sensor nodes is therefore conserved.

In [8][10], spatial correlations among readings of different sensor nodes are utilized. The basic idea is that a node suppresses its reading to the base station if the reading is identical to its neighboring nodes. In addition, the correlations among different types of readings of the same node are utilized in [4]. With the correlations, a low-priced sensor acquisition operation can be substituted for an expensive sensor acquisition to conserve energy for sensor data acquisition. For example, if a particular sensor node's temperature readings are asked, the voltage can be measured to infer the temperature reading because the temperature reading and the voltage reading are highly correlated [4] and measuring the voltage is less expensive.

As it can be seen, many opportunities have been exploited to reduce the cost of monitoring and reporting a group of sensor readings. In this paper we study using temporal periodical patterns to improve the performance of continuous data collections in a sensor network.

In many scenarios, sensor readings exhibit temporal periodicity. For example, the light readings of a sensor node often show daily cycles, increasing at the beginning of a day and decreasing at the end of the day. As an example, Fig. 1 shows a short span of the temperature readings of Berkeley LabData [5]. We can see that the readings exhibit a period with a length of twenty-four hours, where the value of the readings increases from a sunrise and decreases from the sunset.

Such temporal periodicity provides opportunities for conserving the energy of sensor network applications. The basic intuition of using the periodicity is that, as sensor readings are expected to have cyclic behavior, sensor readings can then be derived from past sensor readings. In the following, we use an example to further highlight the

1521-9097/09 $26.00 © 2009 IEEE
DOI 10.1109/ICPADS.2009.58

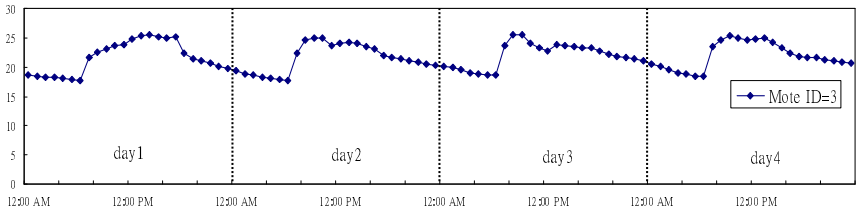


Figure 1. A Temporal Periodicity

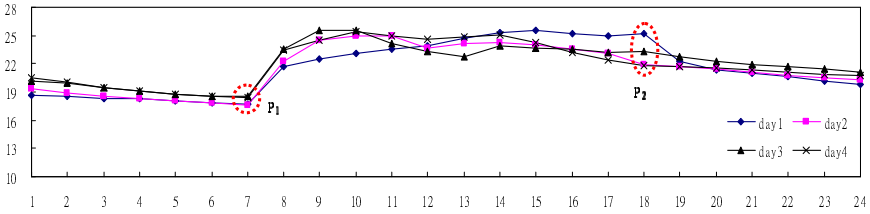


Figure 2. Opportunities for Reducing Cost for Sensor Data Acquisition

intuition and the challenges of using a temporal periodicity for energy conservation.

In Fig. 1, the time series formed by the sensor readings exhibits a period with a length of twenty-four hours. One naïve approach for using the periodicity is to find a representative pattern and use the pattern to predict future readings by assuming that the following days have the same behavior as the representative pattern. However, in general, the perfect periodicity does not exist. As an example, we can break the time series into segments of a period of twenty-four hours (denoted by day1, day2, day3, and day4) and overlap these segments as shown in Fig. 2. We see that there is no perfect periodicity. In Fig. 2, only the readings at some time points show cyclic behavior, e.g. the readings at time point 7, and some points do not, e.g. the readings at time point 18.

Although in general perfect periodicity does not exist, the readings at some time points may show cyclic behaviors, which still provide opportunities for the energy conservation of sensor applications. Our basic idea is to perform sensor data acquisition at the time points that the sensor readings to be acquired are unpredictable, and derive the readings at the time points showing cyclic behaviors from past sensor readings.

In this paper, we propose a novel approach named *PSDA* (*Periodicity-based Sensor Data Acquisition*) to exploit the above stated opportunities for the energy conservation. In the PSDA approach, predicting models for time points in a period are learned from historical readings. A predicting model for a time point determines the *predictability* of the readings to be acquired at the time point. With the predicting models, the future sensor readings can be classified into predictable and unpredictable ones, and sensor data acquisition are needed only when the sensor readings to be acquired are unpredictable. Therefore, the energy for sensor data acquisition and the associated radio communications can be conserved.

To enable the use of the PSDA approach, several challenges are needed to be addressed. First, how do we determine the predictability of the time points in a period? Second, how does our approach work along with streaming sensor readings? Third, how is the accuracy of the proposed approach guaranteed? We identify and address these challenges in this paper.

The contribution of this paper can be summarized as follows. First, we propose a novel energy-conserving approach for continuous data collections in a sensor network. Our approach builds on the observation that the values of the collected sensor data exhibit periodic patterns over time. Second, we provide theoretical analyses for the proposed approach, and show that a tight bound of the accuracy of the reported value can be guaranteed. Finally, comprehensive experiments are conducted to validate the proposed approach. The experiments performed with real data set and synthetic data set validate the effectiveness and efficiency of our approach.

The rest of this paper is organized as follows. In Section 2, we discuss the related work. Section 3 introduces the approach and provides theoretical analysis. Section 4 provides the experimental evaluations and Section 5 concludes this paper and presents some future work.

II. Related Work

The problem of reducing the cost of monitoring and reporting a group of sensor readings have been studied in several fronts [1][2][3][4][6][9][10][14].

Model-based Suppression One prominent direction for reducing the cost of continuous data collection is the model-based suppression approaches [1][3][6][8][14]. In model-based suppression approaches, probabilistic models are synchronously maintained between sensor nodes and the base station. The probabilistic models provide statistical information, such as data distributions, about sensor readings and can be used to predict future sensor

readings. In model-based approaches a sensor node sends its reading to the base station only if the reading differs from the reading predicted by the model maintained in the base station. The base station assumes the readings of the nodes without reporting remain unchanged, and therefore conserve the energy of sensor nodes.

Many well-developed models [1][3][6][14] for fitting the behaviors of individual sensor nodes or a group of sensor nodes have been proposed. In [3][4], time-varying multivariate Gaussians models are proposed to capture the information of sensor readings, including the distributions of individual readings, the spatial correlation among readings from different nodes, and the correlations among different types of sensor readings. In [6], Jain et al. study the use of Kalman filters for fitting the behavior of individual sensor readings, in which Kalman filters between sensor nodes and the base station are kept in synch to reduce the communication cost. In [14], Tulone et al. study the use of autoregressive model for predicting readings at sensor nodes, and exploit the data similarity [14] between sensor nodes that are geographically nearby to further conserve the energy for sensor network.

Deligiannakis et al. [2] propose a novel approach for continuous sensor data collections. The idea is to buffer a series of sensor readings at a sensor node and then extract *critical parameters* from buffered sensor readings. The critical parameters are those can be used to estimate/recovery the readings buffered at the sensor node. In the approach, only the critical parameters are sent to the base station.

The drawback of the model-based suppression approaches is that the existing models such as the multivariate Gaussians models or the autoregressive models are expensive to build and maintain in a large sensor network deployment. In addition, maintaining sophisticated models in sensor nodes is also impractical for the resource limited sensor nodes.

In comparison, our approach only requires some basic operation, e.g., mean and variance computation, for the model construction and maintenance. Moreover, our approach only requires sensor nodes to keep very light-weight information for using the constructed models.

Value-based Suppression Another direction for reducing the cost of continuous sensor data collection is the value-based suppression approaches [8][9][10]. In the value-based suppression approaches, rather than relying on the pre-constructed models for suppressing sensor readings, the suppressions are based on current sensor readings in a sensor network.

In [9], an energy-efficient querying framework is built based on the observation that sensor readings often change infrequently. The basic idea behind the approach is that a sensor node sends its reading only when the reading significantly differs from the last reading.

Madden et al. [8] study the observation that the readings of sensor nodes often are similar to their neighbors' readings. In the study, a snooping technique is developed, with which sensor nodes are allowed to listen the readings their neighboring nodes report. With the snooping technique, sensor nodes suppress their local readings if their readings are identical to the neighbors' readings.

In addition to individual using the temporal suppression and the spatial suppression, the problem of combining temporal correlations and spatial correlations for maximal benefit is further studied in [10].

The value-based suppression approaches reduce the amount of communication. However, the approaches still require sensor nodes to sense to ascertain whether a sensor reading is needed to be sent. In other words, the sensing cost cannot be conserved by the value-based suppression approaches. However, as the applications of sensor networks continue to expand, we find some types of advanced sensor nodes, such as sap flux sensors in [11] or chlorine density sensors, which consume more energy for sensor data acquisition than for communications. This fact makes the value-based suppression approaches not good to be used in these applications.

III. Periodicity-based Sensor Data Acquisition

In this section, we introduce the PSDA approach. In Section III.*A*, we discuss how we construct models for predicting sensor readings, and then in Section III.*B*, we introduce how we use the constructed models.

Before discussing the PSDA, we first provide the assumptions we use through the paper and the problem we want to solve.

Assumptions We consider a sensor network as one which consists of a set of sensor nodes and a base station which has no energy and memory limitations. The sensor nodes are well-synchronized. The base station keeps the network topology and there are no communication delays in the sensor network. Also, in this paper, we assume that sensor devices are reliable, i.e., there are no data acquisition error and communication failures in sensor networks. For the consideration of the failure problems, the readers can refer to [12].

Problem Statement Given (i) a sensor network consisting of a set of m sensor nodes $\{ N_i \mid 1 \le i \le m \}$ and each node N_i produces a value $v_{i,t}$ at timestamp t, and (ii) a data collection task with an accuracy guarantee (ε, δ), where $0 \le \delta \le 1$, $\varepsilon > 0$, which restricts that the maximum absolute error in reporting $v_{i,t}$ must be within the interval $[-\varepsilon, \varepsilon]$ with a confidence probability δ.

Formally, $\Pr(|\hat{v}_{i,t} - v_{i,t}| < \varepsilon) > \delta$, where $\hat{v}_{i,t}$ is the estimate of real value $v_{i,t}$

As an example, the data collection task for temperature readings with the accuracy guarantee ($\varepsilon = 1^{o}$C, $\delta = 0.8$) will report an estimate $\hat{v}_{i,t}$ of the temperature value $v_{i,t}$ with

a confidence probability 0.8 that $|\hat{v}_{i,t} - v_{i,t}| < 1^{\circ}C$. In the following discussion, we call the estimate $\hat{v}_{i,t}$ with the (ε, δ) accuracy guarantee as (ε, δ)-approximation for $v_{i,t}$.

A. PSDA Model

The first step of the PSDA model construction is to collect consecutive readings from a sensor node. The collected readings are used as the historical data from which the periodicity and the uncertainties in estimating sensor readings are learned. In the PSDA approach, the base station maintains a sliding window for each sensor node. The sliding window for a sensor node keeps n consecutive sensor readings of the sensor node.

Given a time series T with length n, the PSDA model construction begins by finding a period in the sensor readings. To this purpose, we use the *Shift-and-Compare* strategy [13], which had been commonly used in periodicity detection applications. As it is named, the basic idea of the Shift-and-Compare strategy is to shift the given time series and then computes the distance between the original time series and the shifted series. Therefore, given a time series with length n, we find periods by iteratively shifting i time units, for i = 1, ... $n/2$, of the time series and computing the corresponding distances. The shifting with minimum distances is output as the period length.

The idea behind the PSDA approach is to only perform data acquisition at the time points the readings to be acquired are unpredictable. For the time points showing cyclic behavior, we derive their readings from past data.

To enable this idea, we have to distinguish the *predictabilities* of the time points in a period. Note that a time point is said to be predictable only if the reading can be estimated with the specified accuracy guarantee. In the following, we introduce the concept of the PSDA model and some of its properties.

Given a historical data T of length n and with a period length p, the PSDA model for the ith time point ($i \in Z^+, 1 \le i \le p$) in the period is defined by the following components:

- *Mean* μ: the mean of the readings at the $(p \cdot s + i)$ th time points in T, $\forall s \in Z^+, 0 \le s \le n/p$.
- *Variance* υ: the variance of the readings at the $(p \cdot s + i)$ th time points in T, $\forall s \in Z^+, 0 \le s \le n/p$.
- *Predictability-Indicator I:* a boolean value that indicates the predictability of the reading at the time point. This indicator value is computed by the following equation.

$$I = \begin{cases} True, & if \ \upsilon < (1-\delta)\cdot\varepsilon^2 \\ False, & otherwise \end{cases}$$

The basic idea behind the PSDA model is as follows. At every time point of acquiring sensor readings, the associated PSDA model can be consulted. If the PSDA model indicates that the reading to be acquired is predictable, then the mean μ of the PSDA model is used as an estimate for the reading to be acquired. We elaborate more on how to use the PSDA models in Section III.*B*.

The following theorem shows that if the variance component υ of a PSDA model is smaller than $(1-\delta)\cdot\varepsilon^2$, the mean component of the PSDA model is an (ε, δ)-approximation for the reading to be acquired.

For ease of presentation, in the following discussion, we use $M_i(\mu_i, \upsilon_i, I_i)$ to denote the PSDA model for the ith time point ($i \in Z^+$, $1 \le i \le p$) in a period.

Theorem 1: *Given a PSDA model $M_i(\mu_i, \upsilon_i, I_i)$, if*

$\upsilon_i < (1-\delta)\cdot\varepsilon^2$,

then μ_i is an (ε, δ)-approximation for the readings to be acquired.

Proof:

Let v be the value to be acquired. By definition, if $\Pr(|\mu_i - v| < \varepsilon) > \delta$, we say μ_i is an (ε, δ)-approximation of v.

We proceed with proof by showing $\Pr(|\mu_i - v| < \varepsilon) > \delta$. First, note that μ_i is the mean of the readings at the ith time point, which can be viewed as the expectation of value v. By the *Chebyshev's Inequality*, we have the following inequality

$$\Pr(|\mu_i - v| < \varepsilon) > 1 - \frac{\upsilon_i}{\varepsilon^2}.$$

By rewriting the given condition, i.e. $\upsilon_i < (1-\delta)\cdot\varepsilon^2$, to $1 - \upsilon_i/\varepsilon^2 > \delta$, and substitutes it into the above inequality, we obtain $\Pr(|\mu_i - v| < \varepsilon) > \delta$. ■

Algorithm Description Given a time series T with length n and a period length p, the algorithm for the PSDA model construction is as follows.

First, we break T into N equal-length segments T_1, …, T_i, …, T_N, each of a length p, where $1 \le i \le N$. We denote the value at the jth time point of T_i as $T_{i,j}$, where $1 \le j \le p$.

Second, for all j (i.e., for all the jth time point in the segments), create a set denoted by $PSDA_j$ that consists of all values at the jth time unit of T_i, $1 \le i \le N$. Formally, we have $PSDA_j = \{ T_{i,j} \}$, $1 \le i \le N$.

Third, compute the means, the variances, and the predictability-indicators for all $PSDA_j$ and output the PSDA models $M_i(\mu_i, \upsilon_i, I_i)$.

We use an example to illustrate the process of the PSDA model construction.

Example 1 Assume we are given the readings in Fig. 3 for the PSDA model construction. For ease of illustration, assume the period length of the time series is given to be eight. The PSDA model construction proceeds as follows. First, we break the time series into four segments, T_1, T_2, T_3, T_4, each with a length of eight time points. Then, we create eight sets, denoted by $PSDA_1$, …, $PSDA_8$, which consist of the values at the jth time point of T_i, for $1 \le i \le 4$ and $1 \le j \le 8$. For example, $PSDA_1$ = {18.5, 18.9, 19.9,

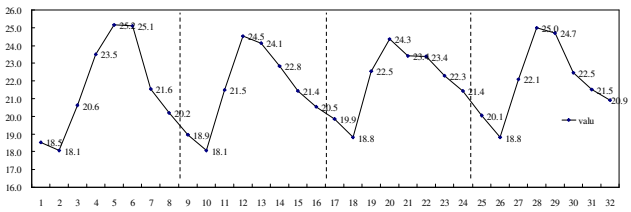
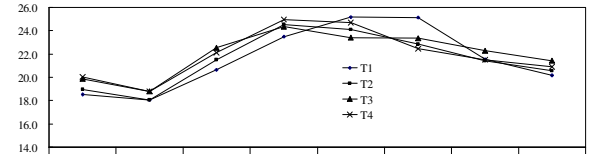


Figure 3(a). Historical data



	PSDA1	PSDA2	PSDA3	PSDA4	PSDA5	PSDA6	PSDA7	PSDA8
T1	18.5	18.1	20.6	23.5	25.2	25.1	21.6	20.2
T2	18.9	18.1	21.5	24.5	24.1	22.8	21.4	20.5
T3	19.9	18.8	22.5	24.3	23.4	23.4	22.3	21.4
T4	20.1	18.8	22.1	25.0	24.7	22.5	21.5	20.9
Mean	19.34	18.44	21.69	24.34	24.35	23.45	21.70	20.77
Variance	0.53	0.18	0.69	0.38	0.58	1.38	0.16	0.28

Figure 3(b). PSDA models

20.1}. Given the accuracy guarantee (ε = 1, δ = 0.8), we have the following PSDA models. They are M_1(19.34, 0.53, False), M_2(18.44, 0.18, True), M_3(21.69, 0.69, False), M_4(24.34, 0.38, False), M_5(24.35, 0.58, False), M_6(23.45, 1.38, False), M_7(21.70, 0.16, True), and M_8(20.77, 0.28, False). Note that the value for judging the predictability, i.e. $(1-\delta)\cdot\varepsilon^2$, is 0.2. ■

B. PSDA Reporting Scheme

In this section we introduce the reporting scheme based on the PSDA models.

The basic idea behind the PSDA reporting scheme is as follows. If a sensor reading at a time point can be predicted, the mean of the associated PSDA model is used as the estimate for the reading to be acquired. Otherwise, sensor data acquisition is performed.

Therefore, when PSDA models for sensor node N_i are constructed, a naïve reporting scheme is to send a sequence of the predictability-indicators $I_1, ..., I_p$ to N_i. Sensor node N_i then performs data acquisition operations according to the sequence.

However, the problem with this naïve reporting scheme is that it may be unaware of the model changes, because the readings not acquired are never known and the associated PSDA models are never updated.

Therefore, we introduce the maximal delay tolerance parameter $\tau, \tau \in Z^+$, which is specified by users to indicate the maximal delay for being aware of anomalies. An anomaly occurs at some time point means that the updates of an associated PSDA model are required. The delay tolerance parameter controls the *freshness* of the PSDA models. A long delay tolerance leads to tardy PSDA model updates and brings more misreports due to the overdue PSDA models.

The PSDA reporting scheme with parameter τ works as follows. On every beginning of a period, the base station computes a *PSDA reporter* for each sensor node and sends the PSDA reporters to the sensor nodes.

A PSDA reporter is a boolean string $S_1, ..., S_p$ indicating whether data acquisition are needed to be performed at time point i. The value of S_i is computed by the following equation.

$$S_i = \begin{cases} True, & if\ I_i = False \\ True\ with\ a\ probability\ 1/\tau, & if\ I_i = True \\ False\ with\ a\ probability\ (1-1/\tau), & if\ I_i = True \end{cases}$$

In computing S_i, one of the following cases occurs.

Case 1: *The associated I_i indicates that the acquisition value cannot be answered under the (ε, δ) accuracy constraint.* In this case, the sensor node is planned for performing data acquisition at time point i in the period and sends the acquired reading back to the base station.

Case 2: *The associated I_i indicates that the acquisition value can be answered with the (ε, δ) accuracy guarantee.* In this case, the sensor node is planned for performing data acquisition at time point i with a probability $1/\tau$. Theorem 2 shows that a sensor node performing data acquisition with a probability $1/\tau$ is sufficient to satisfy the specified tolerance τ for anomaly detection.

Theorem 2: *When the PSDA reporter indicates that a data acquisition operation can be skipped, performing data acquisition with a probability $1/\tau$ is sufficient to satisfy the specified tolerance τ for anomaly detection.*

Proof:

By the PSDA reporting scheme, if an anomaly occurs, we have a probability $1/\tau$ to be aware of the anomaly. Because an anomaly will not vanish if no update is performed, we have a probability $1/\tau$ to detect the anomaly in the next period. Therefore, the expected delay for the anomaly detection can be given by

$$\sum_{i=1}^{\infty} i*(1/\tau)*(1-1/\tau)^{i-1} = 1/(1/\tau) = \tau.$$

That is, performing data acquisition with a probability $1/\tau$ is sufficient to satisfy the specified tolerance τ for the anomaly detection. ■

With Temporal Suppression The PSDA reporter is enhanced with the temporal suppression to further reduce the amount of data transmission. The basic idea behind the temporal suppression is that a sensor node keeps its last transmitted reading v_{last} and only transmits its acquired reading v if $|v - v_{last}| > \varepsilon_2$. The base station assumes that any unreported reading changes within the tolerance ε_2.

There are two things to note about the PSDA reporters with the temporal suppression. First, there are two kinds of unreported sensor readings: the readings not acquired

(by the PSDA approach) and the reading acquired but not transmitted (by temporal suppression technique). Note that there are no ambiguities between the two kinds of readings, because the base station always knows when the real data acquisition are performed (by the PSDA reporter $S_1, \ldots, S_p$).

Second, the error of the PSDA approach and the error of the temporal suppression will accumulate. To enable the use of the temporal suppression, we portion out the original error constraint ε into ε_1 and ε_2, where $\varepsilon = \varepsilon_1 + \varepsilon_2$. We use ε_1 for the PSDA model construction and use ε_2 for the temporal suppression. In the experiments, we study different values for these parameters and discuss the influences of the parameters in more depth.

IV. Performance Evaluation

A. *Experiment Setup*

In the experiments, we perform the performance evaluation with LabData [5]. The LabData records the readings of 54 sensors deployed in the Intel Research Berkeley Laboratory, in which sensor nodes take light, temperature, and humidity measurements once every 31 second from February 28th to April 5th, 2004. We extract the light readings, ranging from 0-800 Lux, of node 2, node 4, and node 19 in the data as test data on which to evaluate our approach. We use the data of the first ten days as initial training data, and run the PSDA approach on the rest data.

Note that there are missing readings in the LabData. We deal with a missing reading of a node by averaging the readings of the previous and the subsequent rounds. In all experiments, we show average values of 100 runs. We evaluate the performance of the PSDA approach by the following two measures: *Error Ratio*: To measure the effectiveness of the PSDA approach, we compute the average number of misreports by the PSDA approach, i.e., how many reported values are away from the actual values than the specified error tolerance. *Energy Saving Ratio*: We also measure the efficiency of using the PSDA approach. We compute the averaged energy conservation with the approach continually acquire and report readings.

B. *Comparisons*

There are six parameters related to the effectiveness and the efficiency of the PSDA approach. They are the PSDA error tolerance ε_1, the temporal tolerance ε_2, the confidence guarantee δ, the sliding window size n, the time between two consecutive readings Γ, (i.e. take a reading once every Γ time units), and the delay tolerance τ for the anomaly detection. In the following experiments, we study the performance of the PSDA approach by varying these parameters. The default values of all parameters used in the experiments are shown in Table I.

Note that in LabData [5] the original value of Γ is 31 seconds. For measuring the performance for different values of Γ, we acquire the value of the sensor readings every Γ/31s time units. For example, if Γ is set to two minutes, then the readings of 1st time unit, 5th time unit, 9th time unit, etc. in LabData [5] are used as the acquired values taken by every two minutes.

Table I. the parameters and their default values

Parameter	Default Value
PSDA Tolerance ε_1 Temporal Tolerance ε_2	ε_1 = 60 Lux ε_2 =15 Lux
Confidence Guarantee δ	δ = 0.8
Sliding Window Size	n= 1440
Delay Tolerance τ	τ = 10 periods
Time between two consecutive readings Γ	Γ = ten minutes

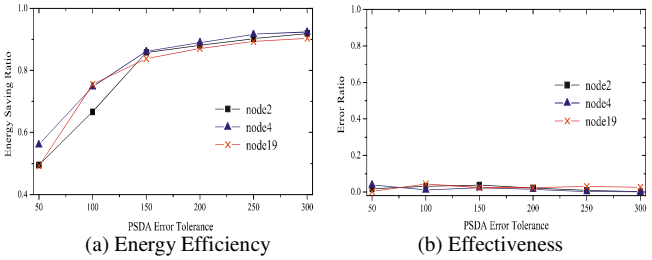
Effect of PSDA Error Tolerance In the first set of experiments, we measure the effect of the PSDA error tolerance by varying the value of the PSDA error tolerance. Fig. 4 shows the results over the default setting, where x-axis is the error tolerance, y-axis in Fig. 4(a) is the energy saving ratio, and y-axis in Fig. 4(b) is the error ratio.

In Fig. 4(a), as expected, we see that higher PSDA tolerances cause higher energy efficiency. The energy efficiency comes from that most sensor readings can be predicted. That is, most sensor readings can be estimated under tolerances without real data acquisition, and therefore the energy for sensor nodes is conserved. In addition, we can see that the PSDA approach provides significant improvements (about 60% energy saving) over a reasonable error tolerance (75 Lux, which is about 10% in proportion to the Lux value domain, i.e. 0-800).

Furthermore, Fig. 4(b) shows the effectiveness of the PSDA approach. We see that the PSDA approach provides very low error ratios (less than 0.05) with respect to all specified PSDA error tolerances. The reason for this effectiveness comes from (i) the confidence parameter setting (δ = 0.8), which bounds the number of misreports reported by the PSDA approach, and (ii) the sensor readings in LabData possess near perfect periodicity, which brings the PSDA approach into full play.

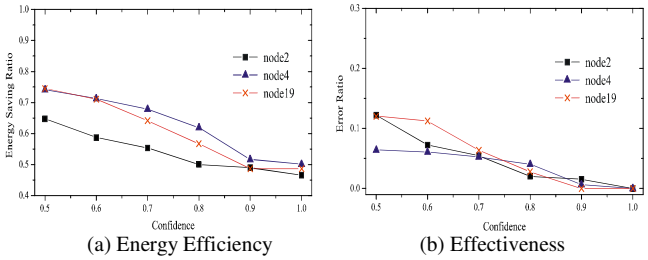
Effect of Confidence In this experiment, we measure the effect of the value of the confidence guarantee δ. Fig. 5 shows our experiment results, where x-axis is the confidence guarantee, y-axis in Fig. 5(a) is the energy saving ratio, and y-axis in Fig. 5(b) is the error ratio. In Fig. 5(a), we see that a lower confidence brings higher energy efficiency, but comes with a high number of misreports, as shown in Fig. 5(b).

Effect of Delay Tolerance In this experiment, we study the effect of the value of the delay tolerance. Fig. 6 shows our experiment results, where x-axis is the delay tolerance, y-axis in Fig. 6(a) is the energy saving ratio, and y-axis in Fig. 6(b) is the error ratio. As discussed, the delay toler-



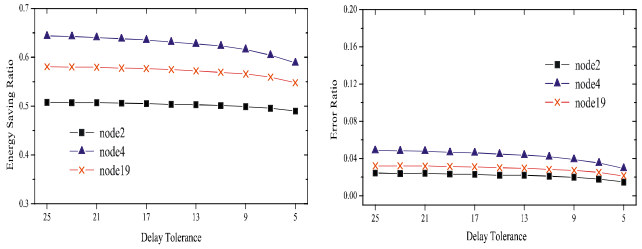
(a) Energy Efficiency (b) Effectiveness

Figure 4. Effect of Error Tolerance



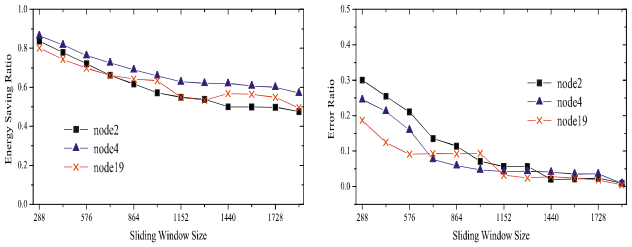
(a) Energy Efficiency (b) Effectiveness

Figure 5. Effect of Confidence



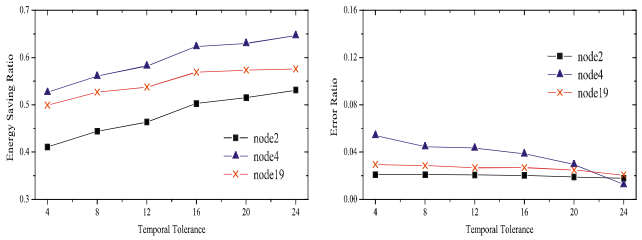
(a) Energy Efficiency (b) Effectiveness

Figure 6. Effect of Delay Tolerance



(a) Energy Efficiency (b) Effectiveness

Figure 7. Effect of Sliding Window Size



(a) Energy Efficiency (b) Effectiveness

Figure 8. Effect of Temporal Tolerance

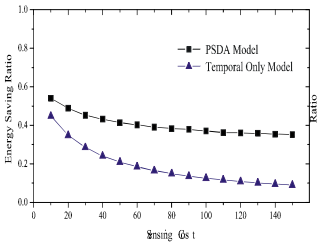


Figure 9. Effect of S. Cost

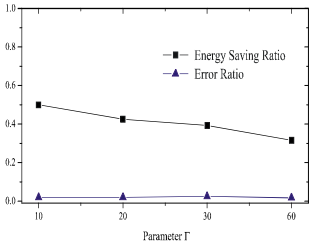


Figure 10. Effect of Γ

ance controls the maximal delay tolerance for detecting the need of model updates. Therefore, intuitively, a large delay tolerance leads to (i) high energy efficiency as few PSDA model updates are needed, which can be observed from the results shown in Fig. 6(a), and (ii) a high number of misreports due to the tardy awareness of the anomalies. However, in Fig. 6(b) we do not observe this degeneration on the error ratio when increasing the value of delay tolerance. In Fig. 6(b), we see that there are no significant differences in varying the value of the delay tolerance. The reason for such results comes from the fact that LabData possesses strong periodicity, with which few updates are needed and therefore less chances the misreporting occurs.

Effect of Sliding Window Size In this experiment, we study the effect of the size of sliding window. Fig. 7 shows the results, where x-axis is the size of sliding window, y-axis in Fig. 7(a) is the energy saving ratio, and y-axis in Fig. 7(b) is the error ratio.

There are two observations for this experiment. First, we see that smaller sliding window size leads to higher error rate. This is because small sliding window are less representative to produce an accurate PSDA model for future data. The second observation is that the energy efficiency decreases when the sliding window size increases, as shown in Fig. 7(a). The reason for such results comes from the fact that the larger the sliding window size is, the more noises the PSDA model construction encounters. The noises come from the missing readings or the shifting of the periods, both increasing the uncertainty for PSDA models and thus making the PSDA approach to be less energy efficient. This influence can also be observed from Fig. 7(b), which shows that the error ratio decreases when the sliding window size increases. The reason is that if the uncertainty in acquiring sensor readings is high, most readings are obtained from real data acquisition and therefore few errors occur.

Effect of Temporal Tolerance In this experiment, we measure the effect of the temporal suppression tolerance ε_2. Fig. 8 shows our experiment results, where x-axis is the value of the temporal tolerance, y-axis in Fig. 8(a) is the energy saving ratio, and y-axis in Fig. 8(b) is the error ratio. As expected, higher temporal tolerance value leads to higher energy saving ratio, similar to the behavior of the PSDA error tolerance.

There is an interesting phenomenon on the result of Fig. 8(b): the error ratio slightly decreases when the temporal tolerance increases. The reason for this phenomenon is that when the temporal tolerance becomes large, the number of misreport becomes less; only the estimate away from the actual value $\pm(\varepsilon_1 + \varepsilon_2)$ is considered as a misreport.

Effect of Acquisition Cost As mentioned, the existing value-based suppression approaches require sensor nodes to sense to ascertain whether a sensor reading is needed to be sent. In other words, the data acquisition cost cannot be conserved, which makes the existing approach not good to be used in the applications where the data acquisition operation consumes lots of energy.

In this experiment, we compare the performance of the PSDA approach and the suppression-only approach by varying the value of data acquisition cost. Both approaches use an error tolerance ε = 75 Lux. Fig. 9 shows the experiment results, where x-axis is the cost of data acquisition and y-axis is the energy saving ratio. We see that when the data acquisition cost become large, the PSDA approach significantly outperforms the other.

Effect of Other Factors We also measure the effect of parameter Γ. Fig. 10 shows the experiment results, where x-axis is the value for parameter Γ and y-axis is the ratio for energy saving ratio and error ratio. We see that the error ratio is always about 2% no matter how we vary the parameter value. The reason is that the misreport of the PSDA approach is mainly controlled by the error tolerance and the confidence, and is irrelevant to the parameter Γ. The same reason also holds for the behavior of the curve of the energy saving ratio.

V. Conclusion

This paper studies using temporal periodic patterns derived from past sensor readings to improve the performance of continuous data collection for sensor networks. We propose a novel energy-efficient approach for sensor data acquisition. The proposed approach works by finding the uncertainties in acquiring sensor readings and avoiding the acquisition that can be predicted such that the energy of sensor nodes can be conserved. The experiments performed with real data show the effectiveness and the efficiency of our approach.

One interesting direction for future work is to build the PSDA models in a distributed manner, which reduces the cost of learning the uncertainty from acquiring sensor readings. We are also currently developing a mechanism for exploiting spatial correlations to cluster the sensor nodes with similar behaviors. With the clustering mechanism, a sensor node is elected as a representative node and its reading used as a representative reading, upon which the other readings are estimated, to further conserve energy for sensor nodes.

Acknowledgment

This research was partially supported by National Science Council, Republic of China, under Grant No. NSC98-2221-E-004-005-MY3 and Grant No. NSC97-2221-E-004-006-MY3. Moreover, this work is supported by ITRI Grant Project 8352B31200.

References

[1] D. Chu, A. Deshpande, J. M. Hellerstein, and W. Hong, "Approximate Data Collection in Sensor Networks using Probabilistic Models," In Proc. of the Intl. Conf. on Data Engineering, 2006.

[2] A. Deligiannakis, Y. Kotidis, and N. Roussopoulos, "Compressing Historical Information in Sensor Networks," In Proc. of the ACM SIGMOD Conference, 2004.

[3] A. Deshpande, C. Guestrin, S. Madden, J. M. Hellerstein, and W. Hong, "Model-Driven Data Acquisition in Sensor Networks," In Proc. of the Intl. Conf. on Very Large DataBase, 2004.

[4] A. Deshpande, C. Guestrin, W. Hong, and S. Madden, "Exploiting Correlated Attributes in Acquisitional Query Processing," In Proc. of Intl. Conf. on Data Engineering, 2004.

[5] http://berkeley.intel-research.net/labdata/.

[6] A. Jain, E. Y. Chang, and Y. F. Wang, "Adaptive Stream Resource Management using Kalman Filters," In Proc. of the ACM SIGMOD Conference, 2004.

[7] S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong, "TAG: a Tiny Aggregation Service for Ad-Hoc Sensor Network," In Proc. of the Symp. on Operating Systems Design and Implementation, 2002.

[8] S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong, "TinyDB: An Acquisitional Query Processing System for Sensor Networks," ACM Tran. on Database Systems, 30(1), page 122-173, 2005.

[9] M. A. Sharaf, J. Beaver, A. Labrinidis, and P. K. Chrysanthi, "TiNA: A Scheme for Temporal Coherency-Aware In-Network Aggregation," In Proc. of the ACM Workshop on Data and Engineering for wireless and mobile access, 2003.

[10] A. Silberstein, R. Braynard, and J. Yang, "Constraint Chaining: On Energy-Efficient Continuous Monitoring in Sensor Networks," In Proc. of the ACM SIGMOD Conference, 2006.

[11] A. Silberstein and J. Yang, "Many-to-Many Aggregation for Sensor Networks," In Proc. of the IEEE Intl. Conf. on Data Engineering, 2007.

[12] A. Silberstein, G. Puggioni, A. Gelfand, K. Munagala, and J. Yang, "Suppression and Failure in Sensor Networks: a Bayesian approach," In Proc. of the Intl. Conf. on Very Large DataBases, 2007.

[13] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average Magnitude Difference Function Pitch Extractor," IEEE Tran. on Acoustics, Speech, and Signal Processing, vol. 22, page 353-362, 1974.

[14] D. Tulone and S. Madden, "PAQ: Time Series Forecasting for Approximate query Answering in Sensor Networks," In Proc. of the European Workshop on Wireless Sensor Networks, 2006.