Optimizing Parameters of Prioritized Data Reduction in Sensor Networks

Cosmin Dini
University of Haute Alsace
34 rue du Grillenbreit 68008 Colmar - France
France
cosmin.dini@uha.fr

Pascal Lorenz
University of Haute Alsace
34 rue du Grillenbreit 68008 Colmar - France
France
lorenz@ieee.org

***Abstract*— Wireless Sensor Networks are popular and proven useful in various service areas. Energy optimization, processing optimization and storage optimization are the main challenges. While there is a debate for complete or partial data extraction from sensors, having special data process functions and operation primitives proves useful for sensor operating systems. To deal with robustness and reliability, data processing at the network/sensor level satisfies some of the reliability requirements, especially when communications are not operational. There are situations where data reduction is an alternative when storage is no longer available and data is accumulating, especially when some sensor links are not operational. Using predictions and optimized parameters to prioritize data reduction is a solution. In this article, we define special heuristics for data reduction using a set of data processing primitives and special data parameters. We apply these heuristics via a methodology that enables various factors, both internal and external to the sensor, to influence the data aging process and the data reduction operations.**

Keywords – sensor, data management, data sensor storage, data priority optimized parameters, prediction models.

I. INTRODUCTION

Requirements posed by unattended data collections in remote areas become very challenging for traditional network deployments. The main problem is raised by the fact that users might look for full collected data, while effective business models take into consideration a small fraction of it.

Most of the WSNs (Wireless Sensor Networks) also perform the essential functions for data processing; one of the most important, in special cases of uncontrolled link availability, is data reduction under several the constraints driven by the nature of the data, the relevance of the data, the data dependency, and the business model using such data. A sensor on a node captures a time series representing the evolution of a sensed physical variable over space and time. Reducing the amount of data sent throughout the network is a key target for long-term, unattended network sensors. A second target, equally relevant, is defined by unattended networks with unreliable links. In this case, gathered data may be rapidly aging and could exceed the storage availability on a given node. Data reduction mechanisms are used to partially handle these cases [1][2][15].

Unnecessary communication as well as appropriate data reduction techniques can be modeled in the case of physical phenomena with a pre-defined, application-dependent accuracy [15]. If an accepted measurement error is bounded as [-*e*, +*e*], only values exceeding the predicted one by +/-*e* will be considered. Similarly, if the errors of the gathered values are within the bonded interval, data reduction can be further simplified in the context of repeated equal measurements.

In this paper, we present a series of heuristics on predictions used to summarize collected data. These heuristics are based on past experience in parameter variation and on the intended use of the data. At the two extremes of data usage are refinement and discovery. Data refinement approaches collection from a perspective of pure prediction where more data is collected around already confirmed scenarios. Data discovery on the other hand tends to ignore known data value patterns by putting more weight on unpredictable corner case scenarios. The difference between the two situations determines what reduction rules are used and how data importance is computed.

The rest of the paper has the following flow. Section 2 introduces the state of the art concerning data management and predictions in sensor networks. Section 3 revisits the model used to reduce collected data. Section 4 includes heuristics for prediction on data processing. Section 5 concludes and identifies future work.

II. RELATED WORK

In this section, we summarize a data processing model introduced in [1][2] and prediction approaches for data reduction [3].

In the past, the database community pushed different data reduction operators, *e.g.*, aggregation and reduction, with no enough flexibility to handle extracting complete raw sensor readings (i.e., using "SELECT *" queries).

There are two specific needs to perform in-network data processing, *i.e.*, (i) to significantly reduce communications costs (energy), and (ii) to deal with link-down situations. In-network aggregation was proposed in [5][6], while data reduction via wavelets or distributed regression in [7][8]. All these techniques do not provide the desired data granularity, as requested by network users.

Managing data in a storage-centric approach was studied in different approaches, based on a reliable connection [11],

 ISBN: 978-1-61208-100-7

additional buffering [9], or collaborative framework [10]. Details on collaborative storage are provided in [2].

OS primitives acting on recurring and non-recurring data collection have been proposed in [1]. Mainly, compression, thinning, sparsing, grain coursing, and range representation were used to deal with data aging in a pessimistic and optimistic approach. As a note, data deduplication was not considered in the above model. To optimize data reduction, concepts of data units, data importance, and compensation factors were introduced in [2]. Mainly, measurements are partitioned into contiguous intervals; data importance relates to the business semantics, while the compensation factor affects the importance in business computation of a given data unit after it has undergone some type of data reduction. The model considers that there is data that cannot be reduced in any circumstance. A mechanism for data dependency between different data units was presented in [2]. Further division of the data units (leading to more flexibility) was not considered at this point. Associated with the new concepts, the following functions were introduced: interval production function (only for recurring data), default compensation function, data importance function, and data reduction function. Solutions based on data redundancy (leading to more robust deployments and measurements) were not considered.

A data reduction specification use case considering data dependency across data units was presented in [2]. Consequently, appropriate values for data importance can be derived considering the constraints on the data importance computation as pertaining to two categories: (i) internal constraints and (ii) external constraints. External constraints are caused by factors over which input data has no effect. Such factors are data age and inherent interest in the data depending on the exact purpose of the data collection. Internal constraints represent inter- and intra-data dependencies.

A prediction model for approximate data collection is presented in [3]. The techniques are based on probabilistic models (BBQ system, [12]). We apply the prediction models to the framework proposed in [1][2] considering also the approximation scheme providing data compression and prediction [13] and predictive models from [14].

In this paper, we consider the PDR components introduced in [2] (Figure 1) to derive appropriate data reduction considering known correlations (spatial, temporal, etc.) and prediction models.

III. BASIC MODEL FOR OPTIMIZATION

We consider a simplified sensor model introduced in [2] with the following modules:

- Storage Engine (SE)

The SE is concerned with writing data to the node's storage. It makes no judgment as to the relevance or importance of the data itself. It simply follows data collection rules established by the business case and sends them to the node's permanent storage. At this time, there may enough space on the storage device in which case the data is simply recorded, or there isn't enough space at which point some data reduction occurs: either on the incoming data, existing data, or both

- PDR Engine (PDRE)

The PDRE contains all the data reduction rules, which are a direct reflection of the business case. They are not constantly applied, but at specific times and with specific space recovery objectives as dictated by the PDR Controller

- PDR Controller (PDRC)

The controller is responsible for monitoring the state of the available storage, and, if dictated by the business case, triggers the PDRE to perform data reduction operations. Deciding what data to target and how much to reduce it is again subject to the business requirements.

Figure 1 shows the interaction of the components shown above: green represents data flow and red represents control paths.

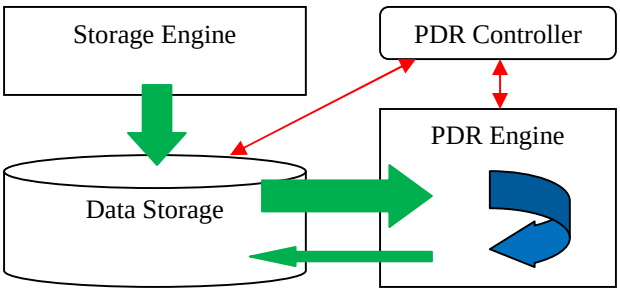

Figure 1. Interaction of main PDR components [2]

Two collection modes are allowed, *i.e.*, Recurring Data Instance (RDI) that represents a series of measurements taken at specified intervals, starting at a given time and ending at a given time, and (ii) Non-Recurring Data Instance (NRDI), representing a single measurement of a parameter at a specific time.

We introduced the following primitive operations that are the basic actions that the PDRE employs to actually decrease the amount of data. Here is a quick review of what they are and how they operate:

- *compression*: a simple data compression algorithm is applied which reduces the amount of space used, but any

 ISBN: 978-1-61208-100-7

further data reduction cannot be applied to the compressed data

- *thinning*: in the case of an RDI, a section of data corresponding to a time interval is discarded
- *sparsing*: in the case of an RDI, for a specific interval, the data sampling rate is decreased and excess data is discarded
- *grain coarsing*: the resolution/precision of a data instance is decreased
- *range representation*: an entire interval of an RDI is replaced by a tuple reflecting on the data that was discarded: minimum value, maximum value, and the average during that interval form the tuple.

To include a complete characterization of the data reduction mechanisms, a few concepts were introduced:

- *Data Units:* Data collections can be both recurring and non recurring. The non recurring collections generate data that is stand alone and considered atomic. It makes sense to consider a non recurring data record as a single data unit (**DU**). Recurring data collections have several values over a potentially long period of time.
- *Data Importance:* Data importance, denoted as **I**, is a value that numerically reflects the relevance of a data unit for the business case.
- *Compensation Factor:* The compensation factor, **K**, is an importance modifier that reflects the data reductions that have already been performed on the specific data unit.

Data importance depends on the business model. In this section, we identify input factors that can be used to establish the importance of a data unit, such as age/collection time, self values, other data units of same instance, and other data instances [2].

The following fucntions are needed to handle the newly introduced concepts for RDI and NRDI

- *interval production function (for RDI only)*
- *default compensation factor*
- *data importance function*
- *data reduction operation function*

In the next section, we present different predictive heuristics for data processing, using the model exposed in Section III.

IV. OPTIMIZATION AND PREDICTION HEURISTICS

Let us assume a deployment of sensors that have the ability to measure the UV Index. The UV Index is a measurement of ultraviolet rays intensity and has a value between 1(low) and 11+(extremely high). The value of this index is collected for the purpose of gaining precise insight into variations during the course of the year.

Expected results are already available:

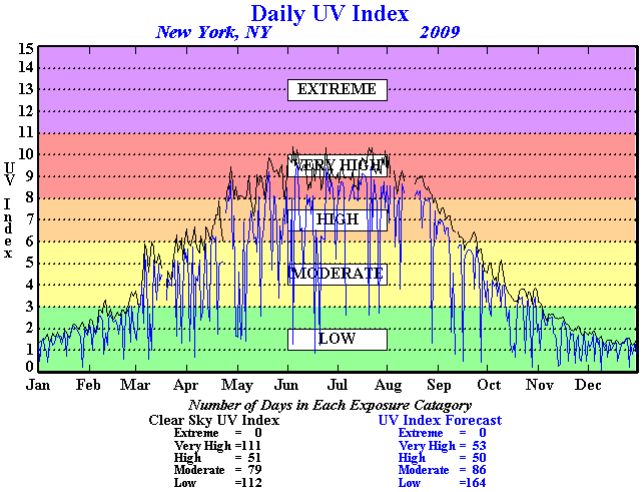

Figure 2. Data UV Index for New York, 2009 [16]

Figure 2 presents the UV Index reading for New York during 2009. We observe that the UV Index has lower values during winter and higher values during summer. There is a natural difference that is expected during a 24 hour cycle, but there are also less obvious dips in the UV Index values. The probable cause for this would be overcast conditions. This naturally interferes with the data collection and is of no interest.

On the other hand, let's consider the example of CO_2 concentration collection in an isolated forest area on a somewhat active volcano. The point of this is not to refine data, but to monitor CO_2 and possibly offer an explanation to unexpected values. CO_2 increase could be caused by a fire in the area or volcanic activity. In such a case, correlations are to be made with seismic sensors, and with temperature and visibility sensors.

In Figure 3 we have a section of interest in a deployment where the objective is not data refinement, but data discovery. We seek correlation between different parameters and possibly seek causality relationships.

 ISBN: 978-1-61208-100-7

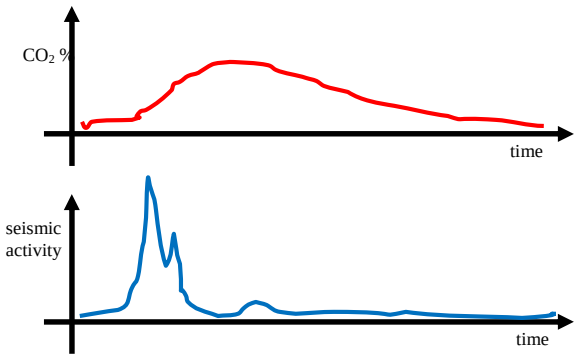

Figure 3. Parameter correlation

The cases presented in Figure 2 and Figure 3 are opposite ends of a spectrum of cases. In Figure 2 we seek to confirm and gain better resolution with respect to expected parameter values, while in Figure 3 we seek to explain observed but not expected parameter values. These cases require different approaches with respect to data reduction.

The data reduction rules that are deployed need to give to the PDR Engine (Figure 1) possibilities to select a reduction approach. Each possibility has different expected outcomes with respect to freed space, effect on the compensation factor, and amount of data importance parameters to be re-calculated.

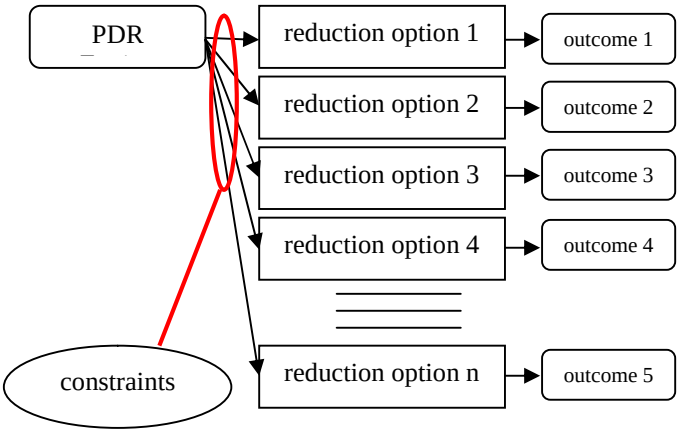

Figure 4. Selecting reduction operation

To investigate optimization of data reduction we define an optimization function as follows:

{fi} = {reduction option 1, . . , reduction option n}

{gj} = {interval production function, compensation factor, data importance, data reduction}

F (x, y) = {fi x gj | constraints}, (1)

where

constraints represent the dependability relations among correlated data.

constraints ::= {context | bounding}

We define the *context constraints* as follows; let *x* and *y* be two variable to gather the values for, and "*oo*" an operator that is defined by

oo = {same, opposite, nil} with the following semantic

x same *y* = when the collection of x is more frequent, then the collection of y should increase too
x opposite *y* = when the collection of x is more frequent, then the collection of y should decrease
x nil *y* = collection of x and y are independent

We define the *bounding constraints* as following:

{x | [-*e*, +*e*], with *e* in *R+*} → p(x/e, y/e') → (2)
{y | [-*e'*, +*e'*], with *e'* in *R+*},

with the semantic: a computation F is not necessary when x's values hold in the bound interval defined by +/-e with the prediction that y's values are bound by +/-e'.

p(x/e, y/e') is the probability that the y's variations hold in the given interval when x's variations are in a given interval; p(x/e, y/e') is derived from the prediction model.

The optimization function triggers the computation of K according to the primitive operations applied for data reduction assuming the prediction model holds. This is an important decision for saving computation power and energy.

These formalisms are used to specify the behavior of the PDR controller mentioned in Figure 1.

We can improve the presented data reduction function considering an average model, i.e., an average ~X of {xi} in the given approximation bounds +/-e predicts a ~Y of {yi} in the +/-e' bounding interval. Computation of ~X and ~Y using in-network data aggregation reduces the consumed computational resources for F, with a reasonable approximation.

Other parameters have to be considered in the data reduction, based on the fact that:

- Some sensing features exhibit typical correlations, *e.g.*, correlation preciseness is inversely proportional with the distance between sensors (temperature)

- The data units can be smaller, especially when working with average values.

 ISBN: 978-1-61208-100-7

Extensive simulations for identifying correlations patterns are needed. Most of the time, in sensor network, and approximate answer is more useful; in special applications, extensive experiments should be conducted to evaluate the performance of data reduction under error threshold define by the bounding constraints. Considering the datasets mentioned in [15], we conclude that a reasonable identification of a correlation pattern requires a sensing period of about one year with a number of samples exceeding 7-8 thousands. For some applications, a tolerated prediction error can relax the bounded interval [$-e$, $+e$].

V. CONCLUSION AND FUTURE WORK

Data reduction in unattended sensor networks with intermittent or non reliable connections is an important computation when considering data storage and data aging.

In this paper, we proposed an optimization function using prediction to map the use of data reduction primitives, optimization parameters (K, I) and dependency constraints (contextual or bounding). The model considers a probability that a variation of a variable is correlated with a variation of another variable. A variant of average values can also be considered.

A simple use case had shown the nature of dependencies and computation challenges for two correlated readings.

Accurate evaluation of the model requires extensive simulations, where combinations of the primitives and data parameters are combined with various type of constraints. We estimate that finding some correlation patterns will favor the use of average values, leading to a reasonable computation effort.

REFERENCES

[1] C. Dini and P. Lorenz, “Primitive Operations for Prioritized Data Reduction in Wireless Sensor Network Nodes”, Proceedings of the 2009 Fourth International Conference on Systems and Networks Communications, September 2009, Porto, Portugal, ICSNC 2009, pp. 274-280

[2] C. Dini and P. Lorenz, “Prioritizing Data Processing in Wireless Sensor Networks”, Proceedings of the 2010 Sixth International Confenrce on Networks and Services, March 2010, Cancun, Mexico, ICNS 2010, pp. 23-31

[3] D. Chu, A. Deshpande, J. M. Hellerstein, and W. Hong, “Approximate Data Collection in Sensor Networks usingProbabilistic Models”, Proceedigns of the 22nd International Conference on Data Engineering. ICDE 2006, Atlanta, USA http://www.cs.umd.edu/~amol/papers/icde06.pdf [Retrieved: August 20, 2010]

[4] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, “Wireless sensor networks for habitat monitoring”, International Workshop on Wireless Sensor Networks and Applications, September 28, 2002, Atlanta, Georgia, WSNA’02, pp. 88-97

[5] C. Guestrin, P. Bodik, T.R., P. Mark, ans S. Madden, “Distributed regression: an efficient framework for modleing sensor network data”, IPSN, 2004

[6] S.Madden, M.J. Franklin, J.M. Hellerstein, and W. Hoing, ”Tag: a tiny aggregation service for ad hoc sensor networks” , SIGOP Oper. Syst. Rev. 36(SI):131-146, 2002

[7] S. Nath, P.B. Gibbons, S. seshan, and Z.R. Anderson, “Synopsis diffusion for robust aggregation in sensor networks”, Proceedings of the 2nd ACM Conference on embedded Networked Sensor Systems, SenSys 2004, Baltimore, MD, USA, 2004

[8] J. Hellerstein and W. Wang, “Optimization of in-network data reduction”, DMSN, 2002

[9] M. I. Khan, W. N. Gansterer, and G. Haring, “In-Network Storage Model for Data Persistence under Congestion in Wireless Sensor Network:, First International Conferenece on Complex, Intelligent and Software Intensive Systems, April, 2007, Viena, Austria, CISIS’07, pp. 221-228

[10] S. Tilak, N. B. Abu-Ghazaleh, and W. Heinzelman, “Collaborative storage management in sensor networks”, International Journal of Ad Hoc and Ubiquitous Computing, Volume 1, Issue 1/2 (November 2005), pp. 47-58

[11] Y. Diao, D. Ganesan, G. Mathur, and P. Shenoy, “Rethinking Data Management for Storage-centric Sensor Networks”, Proceedings of the Third Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar CA, January 7 - 10, 2007.

[12] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong, “Model-driven data acqusition in sensor networks”, VLDB, 2004

[13] I. lazaridis and S. Mehtotra, “Capturign sensor-generated time series with quality guarantees”, ICDE, 2003

[14] S.M. Graham Cormode, M. Garofalakis, and R. Rastogi, “Holistic aggregates in a network world: Distributed tracking of approximate quantiles, SIGMOD, 2005

[15] Y.-Ae¨ . Le Borgne, et al., “Adaptive model selection for time series prediction in wireless sensor networks”, Signal Process. (2007), doi:10.1016/j.sigpro.2007.05.015

[16] National Weather Service Climate Prediction Center [Retrieved; August 28,2010] http://www.cpc.ncep.noaa.gov/products/stratosphere/uv_index/gif_files/jfk_09.png

 ISBN: 978-1-61208-100-7