Multi-resolution Data Communication in Wireless Sensor Networks

Frieder Ganz, Payam Barnaghi, Francois Carrez
Centre for Communication Systems Research
University of Surrey
Guildford, Surrey GU2 7XH
Email: {f.ganz, p.barnaghi, f.carrez}@surrey.ac.uk

Abstract— **There is an increasing trend in using data collected by sensor devices to enable better understanding of the physical world for humans and support the creation of pervasive environments for a wide range of applications in different domains such as smart cities, and intelligent transportation. However, the deluge of data created and communicated and the low-processing capabilities of the used sensor devices lead to bottlenecks in the processing and interpreting of the data. We introduce a data reduction approach that submits high-granular data in times of high activity in the sensor readings and low-granular data in times of low activity. We consider and discuss different methods to measure activity in the data and modify the symbolic aggregate approximation algorithm that uses a fixed window length to adapt the length according to the data activity for ultimately less data communication between sensor node and sink/gateway. We evaluate our approach over real-world data sets and show that reduction of data size while maintaining the features of the data can be achieved.**

I. Introduction

It is predicted that in the next 5-10 years there will be around 50 billion Internet connected devices that will produce 20% of non-video Internet traffic. This data will mainly be created and communicated by sensor devices in wireless sensor networks (WSN) that transmit the data as numerical values. Moreover, most common approaches create and transmit the raw sensor data constantly, even in times where no interesting events occur. Therefore, methods are required to provide multi-resolution data transmission that allows the transmission of high-resolution data (i.e raw data) in times of high activity and aggregated/compressed data in time windows where no events occur. In this paper we propose an approach to create and transmit aggregated patterns of data by applying Symbolic Aggregate Approximation (SAX) [1] to the sensor data unless higher-resolution data is required. We store the raw data in a ring buffer on the local sensor node to provide high-resolution data if required but only transmit aggregated patterns to a gateway node in times of low activity. We define a messaging format (Figure 1) that includes aggregated patterns. These patterns represent the most interesting features and its context information and evaluate traffic consumption on Telos B nodes.

II. Related Work

One approach to reduce the amount of data is data compression. There has been extensive work in the domain of compression algorithms, including widely known de facto standards such as bzip2 [2] and LZE [3]. These common approaches have several disadvantages when applied to the sensor domain. Firstly, the algorithms have been developed for static file compression and not for streaming data. Also they assume that the execution (compression/decompression) of the algorithms takes place on powerful workstations and not processing-limited sensor hardware.

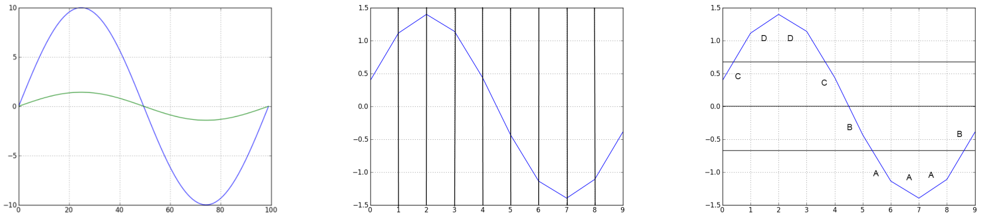
Compression algorithms in the sensor domain, such as the work of Marcelloni *et al.* [4] or others surveyed in [5], use energy and processing -efficient coding techniques to aggregate and reduce the size of data that has to be transmitted. The approaches aggregate and compress all measurements taken by the sensor. In this work we focus on reducing the data that has to be transmitted by only submitting data when there is a certain activity in the data.

We use the SAX algorithm that includes a dimension reduction technique to reduce the amount of data. SAX consists of two steps: piecewise approximation aggregation (PAA) transformation and the transformation of the numerical data into a string representation. SAX creates a compressed symbolic representation of time series data by taking the average of data windows. The averaged PAA data is then split vertically. The break lines are distributed vertically according to the Gaussian distribution. Each window is assigned a letter, depending on which break line the average of the window resides under. This process is depicted in Figure 2. The main advantages of SAX are low processing costs and high data reduction while retaining the main features. SAX leads to dimensionality and numerosity reduction and is the building block for many pattern and outlier detection algorithms ([6], [7], [8], [9]). The computation of a SAX word involves the process of reducing the dimensionality of a time-series by averaging the data; this is shown in the following: SAX transforms a time-series X of length n into a reduced vector $\bar{X} = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_m)$ with length m. Each element $\bar{x}_i$ is calculated with the formula shown in equation 1.

$$\bar{x}_i = \frac{m}{n} \sum_{j=n/m(i-1)+1}^{(n/m)i} x_j \tag{1}$$

By applying the process to the series $X = (4, 8, 3, 2, 1, 1, 1, 1, 1, 1, 10, 5)$ with length $n = 12$ and

Fig. 1: Dimensionality Reduction Process of SAX



(a) The blue line represents 100 data points of a sine curve. The green (dotted) line is the curve after normalisation

(b) The normalised curve is divided into 9 windows. Each window represents the mean of 100/9 datapoints from the original data. That leads to a compression rate of m/n 9/100

(c) The output windows of the PAA is then divided vertical. Each segment is assigned to a letter. The segment in which the curve is in per window forms the SAX word, in this case CDDCBAAAB

an aimed reduced vector $\bar{X}$ of length $m = 6$ we get the result $\bar{X} = (6, 2.5, 1, 1, 1, 7.5)$.
Subsequently, the data gets transformed into a symbolic representation, according to the break lines defined by the Gaussian distribution shown in Figure 1. In this example the output vector is transformed into the symbolic representation $CBAAAC$, this eases the indexing and the comparison between different time-frames, though, this is not the focus of this work.
The drawback of the initial SAX algorithm is the use of a fixed length of m that leads to the same aggregation level even in the case of low activity in the data. In the aforementioned example the values $(1, 1, 1, 1, 1, 1)$ are reduced to $(1, 1, 1)$ where it could have been aggregated to one value. This is due to the fixed length of m. In our approach we introduce a variable length m to get a smaller reduced vector $\bar{X}$ at times where there is low activity in the data. In the following section we select an appropriate method to reflect the activity in the data and modify SAX for a variable granular selection of m. In our approach we aggregate the sensor data with the help of SAX applied to the data and include it in a multi resolution message depicted in Figure 1.

III. Multi-Resolution Data Communication

The aim of this work is to reduce the amount of data by having a high granular representation of the sensor measurements at times when there is high data activity and a lower granular representation in times of low activity. This requires a message format that stores the period when the observations have taken place and the SAX representation. Due to the differing lengths of the SAX patterns, the observation period is required to reconstruct the original data from data that has variable granularity.
To select the different levels of granularity a method has to be introduced that based on the data activity chooses the right length m of the reduced data. In the following the details of the message format and the variable granularity selection method are described.

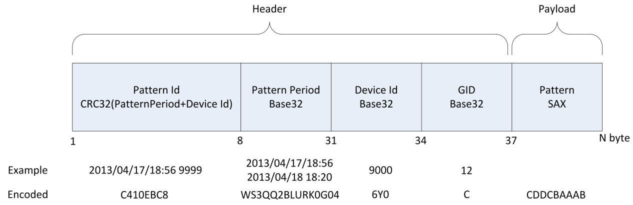


Fig. 2: Multi-Resolution Message Format showing the message structure and sample message in readable data and encoded.

A. Multi-Resolution Message

The message includes the pattern id, obtained by CRC32 hashing the start and end time of the window and the sensor device id, the base32 representations of the pattern period, device id and gateway id, and as the payload the SAX representation over that period. In case higher resolution data (i.e raw data, or a higher SAX resolution) are required, the raw data is stored in the local cache of each sensors and can also be transmitted. The node can be queried with the pattern id from the multi-resolution message to obtain more precise values.

B. Variable Granularity Selection

In order to choose the right granularity and therefore estimate the length of the reduced vector m, the correct method to measure the activity in the data has to be selected. In the case of high data activity a finer granular representation is desired, thus m has to be close or equal to n. In the case of low or no activity, m has to be low or close to 1.
To measure the activity in the data we pre-selected four statistical methods that can give insights about the activity in the data, namely variability measured as variance var, maximum max, minimum min and mean $mean$. The particular selection of m according to the different methods is described and discussed in terms of its applicability as follows:

1) max: Firstly, an average maximum of historical data is identified. If the currently observed data frame is close

Fig. 3: Granularity selection based on the quartiles of the data distribution for the variance

$$f(var) = \begin{cases} m_{low}, & \text{if } var \leq Q_1 \\ m_{medium}, & \text{if } var \leq Q_2 \\ m_{high}, & \text{if } var \leq Q_3 \\ n, & \text{otherwise} \end{cases}$$

to or higher than the previously identified maximum m is chosen closer to n to achieve the desired high granularity. However, the application of this method is only useful for data that has interesting outliers higher over a certain threshold; for example, this could be applied to presence data where presence could be identified using local maxima.

2) min: Selecting m based on the minimum has the same applications as choosing the maximum value discussed above; however it is applicable where a higher granularity should be achieved for small values.
3) $mean$: Taking the mean to select the granularity will result in a higher granularity data values that are stationary around a certain value. This reduces the granularity in cases where there are many outliers.
4) var: The variability measure defines how far values are spread out. This can be used to create a higher granularity in values that are more distant to the mean of the data. This includes the features of the min, max approaches. However, it does not favour values that are around the mean. In this work, we assume that the values away from the mean are more interesting and those values should be represented with a higher granularity then data that is close to the mean.

To select m, we introduce functions for each statistical method that lead to a higher granularity based on the distribution of the data. In figure 3 the function to select the granularity and thus m according to the variance is shown. In the case that the variance is in the first quartile of the distribution a smaller m is selected. If the variance is within in the range of second quartile,then a medium m is selected.

IV. Evaluation

Before we evaluate our approach we show that changing the output length m has an impact on the data size and the data reconstruction. We use several data sources, namely presence, power consumption, light level and noise level measured in an office environment over 6 weeks resulting in 280000 samples. We divide the dataset into 50 frames with a length of 560 samples. We show in the top graph in figure 4. how the correlation between original data and reconstructed data created by using a length m from 1 (lowest granularity/frame) to 560(using the full frame). In the bottom graph the data size under the different lengths m is depicted.

We evaluated our approach in two steps. First, we study the extension of SAX with a variable output length m using different granularity selection methods as mentioned in section III.b The data reduction size and the Pearson correlation between original and reconstructed data from the dimensionality reduction algorithm are used as evaluation metrics.

We then evaluate a communication scenario with the multi-resolution messages introduced in section III.a by measuring the traffic size in a real-world setup on Telos B nodes.

A. Variable Granularity Selection

It is important to select the right function to detect activity in the data and chose the variable length m for the dimensionality reduction algorithm. We use a dataset obtained from a presence (passive infra-red) sensor deployed in an office environment. The dataset contains 55000 data samples measured over one week. The data is transformed into a reduced dataset using the SAX algorithm. However, instead of using a fixed window length n, we choose a variable length m by measuring the data activity. The data activity is measured using the common statistical methods var, $mean$, max and min. After the transformation, we compare the similarity of the original and reconstructed dataset by using Pearson correlation and also compare the size of the original and reconstructed datasets. By choosing the var as selection method, the dataset is reduced by 36% with a correlation factor of 0.94. For $mean$ 27% and 0.95; For max 0.68% and 0.92; And for min 29% and 0.99 respectively. As discussed earlier, does the reduction and reconstruction strongly depend on the underlying dataset and this evaluation of only one dataset can be insufficient. This explains the large reduction and high correlation when choosing the min approach. However, we argue that, as a general approach, the variability of the data, thus outliers in respect to the mean, reflects the range of most interesting values.

In Figure 5, we show the data and the changing granularity of the window length with the variance method. In times of high activity the granularity rises and vice versa in times of low activity.

B. Multi-Resolution Message

We evaluated our approach by simulating data transmission from Telos B nodes running the TinyOS operating system. During the simulation we collected the data from the temperature sensor on the Telos B node every hour for 30 days, leading to 720 samples. Each raw data sample to be transmitted includes timestamp, value, device id and gateway id (where the data is submitted to) with a size of 22 byte per sample and 15.840 bytes during the whole sampling period.

We ran our simulation with different window lengths over the same 720 samples. We created a pattern payload with the multi-granular SAX representation. Our results show that we can reach a reduction in transmission traffic of the data by sending our messages. Instead of transmitting 15.840 bytes we could reduce the size to 10.240 bytes (36%) and a correlation factor of 0.93. However, if more precise data is required, the user/application can query the sensor that caches the raw data.

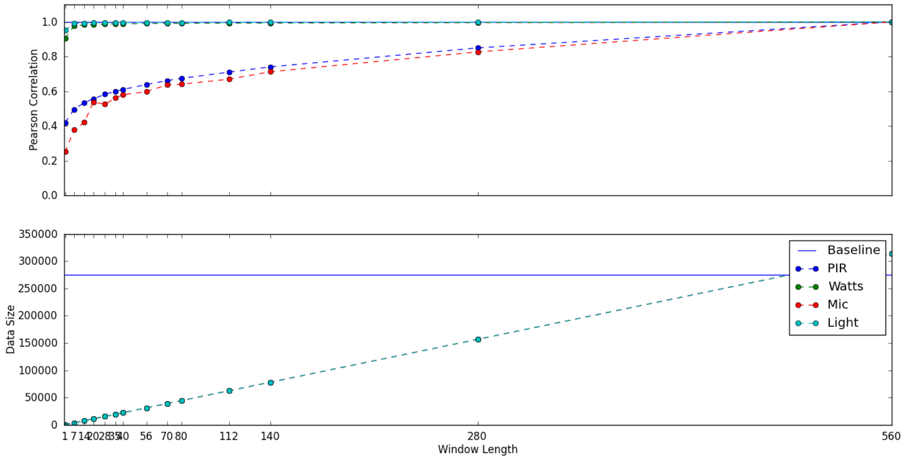


Fig. 4: Impact on data size and correlation of reconstructed data by using different window length m

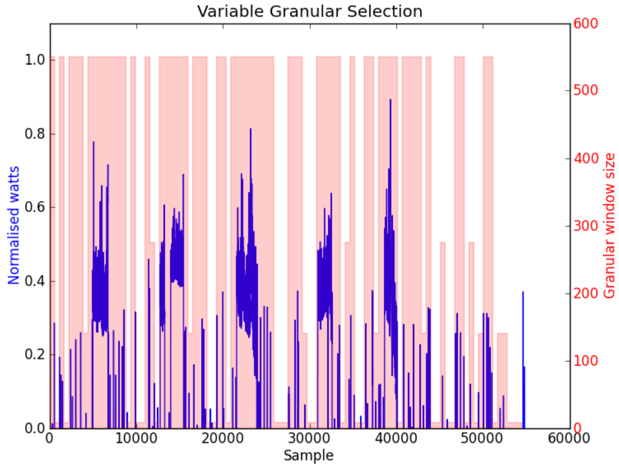


Fig. 5: Change of the window length m (red) over the dataset, showing smaller m for less active areas and higher m for more active areas

V. Conclusions

In this work we introduce a multi-resolution message format that represents data observations aggregated into high granular representations in times of high data activity and coarse granular representation in times of low activity. This leads to a reduction in to traffic from the sensor devices to a sink/gateway node by maintaining high reconstruction rates of the original data. We show that it is important to choose the right method to measure activity in the data, and that it is dependent on the dataset.

However, if the main interest in the later processing is in the outliers of the observed data we recommend selecting the variance as a measurement for data activity.

Future work has to evaluate this statement by taking larger and more diverse data sets into account. As a possible solution to finding the right measurement, statistical tests such as the t-test could be applied to identify similar distributed datasets in order to apply the best activity measurement function. Also, the entropy of data could be taken into account when selecting the length of the output vector based on the entropy level of the data.

Acknowledgment

This paper describes work undertaken in the context of the EU FP7 Citypulse project contract number: 609035.

References

[1] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, ser. DMKD '03. New York, NY, USA: ACM, 2003, pp. 2–11. [Online]. Available: http://doi.acm.org/10.1145/882082.882086

[2] J. Seward, "bzip2 and libbzip2," *avaliable at http://www. bzip. org*, 1996.

[3] G. Seroussi and A. Lempel, "Compression using small dictionaries with applications to network packets," Feb. 14 1995, uS Patent 5,389,922.

[4] F. Marcelloni and M. Vecchio, "A simple algorithm for data compression in wireless sensor networks," *Communications Letters, IEEE*, vol. 12, no. 6, pp. 411–413, 2008.

[5] N. Kimura and S. Latifi, "A survey on data compression in wireless sensor networks," in *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, vol. 2. IEEE, 2005, pp. 8–13.

[6] E. Keogh, J. Lin, and A. Fu, "Hot sax: efficiently finding the most unusual time series subsequence," in *Data Mining, Fifth IEEE International Conference on*, nov. 2005, p. 8 pp.

[7] Q. Yan, S. Xia, and Y. Shi, "An anomaly detection approach based on symbolic similarity," in *Control and Decision Conference (CCDC), 2010 Chinese*, may 2010, pp. 3003 –3008.

[8] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: a novel symbolic representation of time series," *Data Mining and Knowledge Discovery*, vol. 15, pp. 107–144, 2007. [Online]. Available: http://dx.doi.org/10.1007/s10618-007-0064-z

[9] D. Minnen, C. Isbell, M. Essa, and T. Starner, "Detecting subdimensional motifs: An efficient algorithm for generalized multivariate pattern discovery," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, oct. 2007, pp. 601 –606.