11

An Enhanced Spatial Temporal Adaptation Model for Precise AQI Estimation

Abraham Amal Raj B
Research Scholar
Department of Computer Science & Informatics,
Maharishi Arvind University, Jaipur, Rajasthan
Dr. Mahaveer Sain
Associate Professor
Department of Computer Science & Informatics,
MAISM, Jaipur, Rajasthan, India
Dr. Dharmveer Yadav
Assistant Professor,
Department of Computer Science,
St. Xavier's College Jaipur, Rajasthan, India
dharmveeryadav@sxcjpr.edu.in

Abstract

The application and consistency of wireless sensor networks (WSNs) in acquiring the data for AQI (Air Quality Index) fundamentally depend on data redundancy. An inherited value of sensory data is spatial and temporal similarity. Reducing this spatio-temporal data redundancy can save significant node power and bandwidth. To reduce data redundancy, essentially every one of data collection method employs either temporal or spatial correlation. Through analogously based sub clustering, manipulation of spatial and temporal correlations among sensor data was performed on Spatial and temporal correlations using the Advance Multiple Data Prediction Interface (*SAMI*) and in our planned model there is a reduction in the spatial redundancy of sensor data. A single instance node can represent nodes that are carefully related by similarity-based sub-clustering. Using model-based prediction techniques can help reduce temporal redundancy and propagate only a subset of sensor data while predicting termination. This method provides a user-defined error threshold for data while significantly reducing power-consuming traffic. Being a distributed methodology, the planned work is in ascending order. This work resulted in a data reduction of up to 75% for the occasional collected system data kept within 0.6°C error.

***Keywords**:* WSN, AQI, similarity-based clustering, data reduction, prediction

Introduction

With forays into chip integration, MEMS, and RF expertise, WSNs are being used in diverse applications, which include environmental monitoring. They rely on growing an extensive terminology data collection on the WSN to maintain the required level of accuracy. Each sensor node within the WSN acts like an identity-wrapped structure that facilitates sensing, computing, and communication. A restricted energy supply is the main limitation of sensor nodes. Wireless communication, which plays an important role in different aspects depending on the type of acquisition performed, is the most important energy-consuming function. Looking at it another way, computing was considered the least energy-consuming activity. Achieving long life while maintaining minimal power consumption is a key goal of excess data collection when using WSNs, and high resolution and high quality are required to enable meaningful analysis. The quantity of live sensor nodes and the amount of data sent by the sensor nodes determines the cost of distributed monitoring.

One method of collecting periodic data[1] is that the node first collects the environment to get the finest granularity of data. The data is then continuously transmitted for a period of no interest. This transmitted data helps enable highly complex data analysis. If the data is continuously collected and the WSN lifetime is also shortened, this process becomes expensive and sophisticated. If bandwidth is limited, introducing multiple nodes will result in uneven communication, excessive data collisions, traffic congestion, and reduced throughput. Redundant data accounts for a large portion of the total amount of data transferred.[2] Although not informative, redundancy consumes a certain amount of network resources. Advantageously, data transmission can be aggressively reduced to save power without sacrificing a significant reduction in observation reliability. This leads to the presence of spatio-temporal correlations in the sampled data. When sensor nodes are in close proximity to each other, their observations might have similarity, so the values of adjacent nodes can be easily predicted. There will be high spatial correlation among the sensors that are physically nearby. To classify similar nodes with the peak energy of CH formation, *Turan's* theorem was employed which is founded on extremal graph theory. By observing the resemblance in size

[1] M. Li and Y. Liu, "*Underground Coal Mine Monitoring with Wireless Sensor Networks*," ACM Trans. Sensor Networks, vol. 5, no. 2, pp. 1-29, 2009

[2] D. Culler, D. Estrin, and M. Srivastava, "*Overview of sensor networks*," Computer, vol. 37, pp. 41–49, 2004.

and trend of the produced data, it is easy to predict that they are adjacent nodes, as they are also highly spatially correlated. Future results are simply predicted from previous data collected at the same node if consecutive data are more similar than the sampling frequency, at sensor nodes. Approximating signal trends from time correlations can help predict future data.

Planned work uses both inter and intra-sensor temporal and spatial correlations to reduce communication effort without compromising precision. Sensors within a cluster with analogous adherences are clustered into different subclusters. One of the subcluster's sample nodes is chosen to represent the entire subcluster. The Sample node uses LMS filters to accumulate a time correlation model based on previous data collection. The created model corresponds with other subcluster members and CHs. The sample node updates trend changes by updating the appropriate filter coefficients. When a user-defined tolerance value surpasses observed and predicted data, subcluster members begin transferring data. This system of ensuring the accuracy of custom nodes effectively reduces the communication cost of the regular reporting framework.

A prediction system value was performed on a synthetic data set with many correlations. In terms of energy, this system is more efficient. Highest accuracy is achieved with collected data. Improved balancing of node energies within subclusters without negotiating data accuracy. A temporal prediction approach for identifying subcluster heads implemented the idea of large subsets given by *Turan*.

Survey of Related Works

Energy efficient functionality is a crucial matter in any WSN strategy. Most WSN applications are dominated by limited energy as the bottleneck and are based on energy conservation, and much work has been done on energy conservation in WSNs. An important direction for energy conservation in WSNs was published in Anastasi *et al*.[3] discusses where turn-on intervals depend on mobility and data control and describes those methods. Redundant data can be reduced in a data-driven way. Reduce traffic volume, conserve bandwidth, and avoid data collisions to maintain node power.

[3] Anastasi, G., Conti, M., Di Francesco, M., and Passarella, A. 2009. *Energy conservation in wireless sensor networks: A survey*. Ad Hoc Networks 7, 3 (May), 537–568.

By exploiting temporal correlations between consecutive data, temporally redundant data can be reduced in several ways.[4] From recent data history, reduce communication overhead by predicting future data. Applying linear regression methods,[5] some of the data prediction methods could take advantage of temporal correlations among sensory data, but the absence of adaptableness to dynamic fluctuations in the input signal also reduces accuracy. Predict future sensor data from previous data history using ARIMA-based methods.[6] ARIMA needs rich baseline data. This requires intensive computations and deprives the set of predictions when there are many turning points. Prediction was performed using PCA, which elaborated on the previous model characterization. In the planned model, we will employ the LMS algorithm for model-free prediction filters to take advantage of temporal correlation.[7] The predicted methodology's data dynamics are amply adaptable and easy to compute.

PRESTO[8] builds a model of higher-layer proxies that help you correlate the data observed by each lower-layer sensor. When remote sensors start comparing collected data to this predictive model and push data to detect unusual trends, the model's observations deviate from the predicted values. PRESTO only considers temporal correlation and ignores spatial correlation of nearby sensors. To sample data from different sources, various data acquisition methods have been predicted for modulation of the spatial properties of active sensors, and linear models are predicted by capturing spatial correlations. This model enables maximum sensor nodes to switch to sleep mode. A linear combination of datasets from functioning sensor nodes can be used to derive sensor node readings to a certain degree of accuracy. Conversely, in practice,

[4] S. Chatterjea and P. Havinga, "*An Adaptive and Autonomous Sensor Sampling Frequency Control Scheme for Energy-Efficient Data Acquisition in Wireless Sensor Networks*," Proc. IEEE Fourth Int'l Conf. Distributed Computing in Sensor Systems (DCOSS '08), 2008

[5] Carlos.Carvalho, Danielo. G.Gomes, Nazim Agoulmine and José Neuman de Souza, "*Improving Prediction Accuracy for WSN Data Reduction by Applying Multivariate Spatio Temporal Correlation*," Sensors , vol. 11, pp. 10010–10037, Oct.2011.

[6] Li and Wang, "*Automatic ARIMA modeling-based data aggregation scheme in wireless sensor networks,*" EURASIP Journal on Wireless Communications and Networking,2013,2013:85.

[7] Santini S and Römer K, "*An adaptive strategy for quality-based data reduction in wireless sensor networks*," in Conf. networked sensing systems, 2006, pp. 29-36.

[8] Ming Li; Ganesan, D.; Shenoy, P. "*PRESTO: Feedback-Driven Data Management in Sensor Networks*", IEEE/ACM Transactions on Networking (Volume:17, Issue: 4), Aug. 2009, pp.1256 - 1269

many systems would not be linear. Additionally, the method by which ASAP[9] chooses the appropriate work node has not been explained.

Sub-clusters are created by selecting correlated sensor nodes, which then choose a subset of samplers to continuously collect data from within the sub-cluster. For predicting non-sampler datasets spatial correlation can be useful. Because ASAP uses a probabilistic model used only by ASAP to validate forced sampling periods, error bounds on forecast data are not guaranteed. Unusual trends among forced samples can go unnoticed if error prediction is not performed correctly. Sub-clusters for planned work can be built based on closely correlated sensor nodes. Spatial correlation is checked at each round of data collection. All the above methods mainly focus on temporal or spatial correlation.

Combined Dual Forecasting

System Architecture & General Overview

The SAMI architecture's basic functional representations and specifications are outlined, along with a brief explanation of the underlying mechanisms. To minimize data communication within the network, spatial and temporal correlations between sensor data in planned work are utilized. The process begins with the formation of subclusters, where the inference of spatial correlations between sensor data is performed, by grouping strongly correlated sensor data. A sub-cluster is represented by a single sampler node, which eliminates redundant data from neighbouring nodes. LMS filters estimate the temporal correlation of sensor data to predict future data. Because of the predictive data, a specific subset of the data is sent that differs from the desired data. This technique helps filter all spatially and temporally redundant data. The use of ensemble forecasting techniques enables the detection of abnormal trends in sampler nodes, and the propagation of these anomalies to sinks.

This projected system pursues a three-layer structure. The bottom layer has N nodes, randomly distributed across the array. Each node operates as a system that calculates the source of finite energy powering its communication modules. High-energy cluster heads associated with groups of spatially close nodes form node clusters, serving as a secondary layer. The nodes' data is aggregated by the CH and transmitted to the base station. The closely correlated nodes in these clusters are then further divided into subclusters, forming a third layer. Each subcluster

[9] Gedik, B., Ling Liu, Yu, P.S., "*ASAP: An Adaptive Sampling Approach to Data Collection in Sensor Networks,*" IEEE Transactions on Parallel and Distributed Systems, (Volume:18, Issue: 12) Dec. 2007, pp- 1766 – 1783.

has a subcluster header (SCH) that represents it. SCH extrapolates the SCH-generated data to represent the entire subcluster. Figure 1 illustrates a three-tier design.

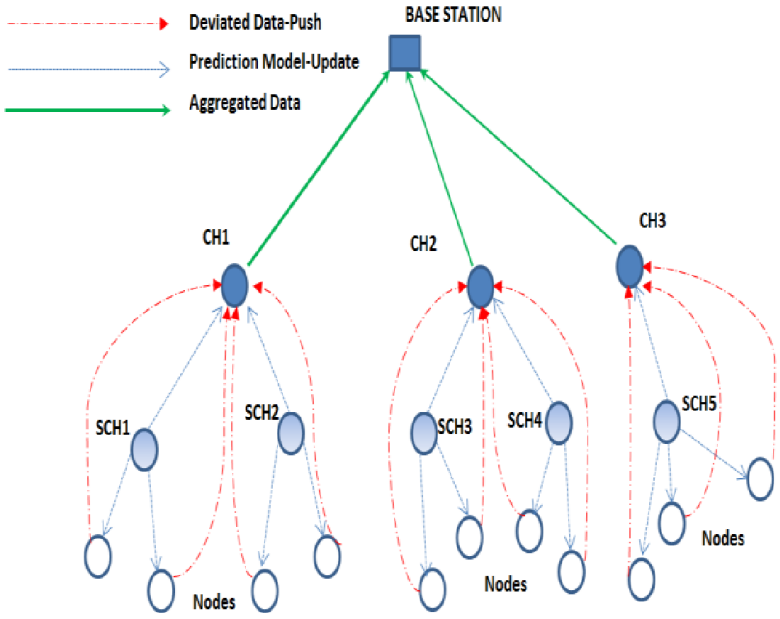

Figure 1 Three-Tier Design of SAMI

A work flow is presented in four segments. The weight-based passive clustering technique is utilized to designate energy nodes as CHs and create groups around them. The cluster head gathers information from its affiliates and sub-divides the cluster further depending on data resemblance. The node with the highest energy represents a sub-cluster head (SCH) and a temporal-correlation prototype is created by the sub-cluster head using LMS-based filters and past observations. This model facilitates future data prediction with custom error tolerance, and is shared by the members of the subclusters and the CH. When the deviance between the forecasted and observed data surpasses time error tolerance, a comparison of these two data sets is made in the fourth phase to update the model, triggered by each random sample. During each sample time, as the deviation of the error threshold grows, the subcluster members compare their forecasted data to their observations and combine the information. The CH predicts data based on the scheme, examines for updates from the Secondary Cluster Head (SCH), and adjusts by verifying inputs from its subcluster members. The maximum size and

spatial correlation configuration are governed by extreme value graph theory based on *Turan's* theorem, as proposed by Tolhuizen *et al.,* using the following properties:

$$ex(v, R) = \{t(E) | R \not\subset E, |S(E)| = v\} \quad \text{.....(a)}$$

If updates are not received from the SCH, then the model is deemed accurate for that moment. The fit takes into account all the data from the subcluster. The subclusters censor transmission from the other members. If communication is detected, the corresponding data set for the equivalent subcluster is replaced.

Passive Clustering for Energy Efficiency

In most Wireless Sensor Networks (WSNs), scalability is accomplished by grouping nodes that are in close proximity and monitoring both route stability and bandwidth maintenance. The computational load is distributed among clusters of CH by making various complex inferences. To perform numerous computations on different data series, the network is divided into clusters of spatially close nodes, led by energy nodes. The CH election algorithm follows a deterministic approach that ensures a consistent distribution of CH. To maximize energy conservation, high-energy nodes are selected as Cluster Heads (CHs). A passive clustering method was developed to reduce power consumption during the clustering process, where the announcement delay is determined by the node's remaining power. As CH has multiple tasks, it experiences high energy depletion, so the selected CH must have ample residual energy. The proposed work uses residual energy as a weighting parameter for CH selection, as residual energy recovery is an internal task that does not require communication.

The process of selecting a CH in passive clustering is based on the principle that the first declaration wins, meaning that the first node to declare itself as CH will become CH. In previous work, the declaration could be delayed, which may result in a low-power node being chosen as CH, decreasing the energy efficiency of the system. When selecting the best node, the declaration lag is inversely related to the node's residual energy. At the end of the declaration lag, the node declares itself as the cluster head. However, if a node receives another declaration before its own declaration delay is over, it will not become the Cluster Head (CH). The latency, T_w, of node n can be calculated as follows:

$$T_w(n) = k/E_{res} \quad \text{.....(b)}$$

Here, the residual energy of node "n" is represented by E_{res}, and "k", a constant. When a node entertains several CH declaration requests from numerous nodes, it selects the node with the highest residual energy as the CH and forms a cluster with it.

Leveraging Spatial Correlation

By taking advantage of the spatial proximity of nodes, the WSN aims to reduce the amount of redundant and spatially similar data transmitted, thus conserving network power. When the WSN is organized into clusters, the Cluster Head (CH) gathers data from all nodes in the cluster, which is then aggregated and transmitted to the sinks. In this clustered approach, the CH filters out any data that is redundant based on spatial proximity, thereby reducing the amount of unnecessary data flow. Still, it may be more efficient to eliminate redundant data at the individual node level, rather than relying on the cluster head to filter it. In planned work, based on the sensor time series, different subclusters are assigned by the cluster leader according to data similarity. SCH nodes report data for each corresponding subcluster to CH each time. As a result, spatial data is inherently sifted if they are redundant.

The formation of subclusters takes place in three steps. First, nodes with high energy levels are identified as the subcluster head (SCH). Then, the nearest neighbouring nodes are determined. Finally, the SCH data row is compared with other neighbouring data rows. Size and trend similarity are compared. The subcluster leading the SCH will add nodes if the neighbours are similar in size and trend. Secondary energy nodes are identified considering other cluster members. During this process, replication runs until all cluster nodes have been added to the subcluster.

In three steps, the formation of subclusters is completed. The nodes with the most residual energy are identified as SCH and their nearest neighbours are then located. The SCH data row is then compared with other adjacent rows. Through SCH, the CH receives data from within each subcluster. Thus, the residual energy of other nodes should be lower than that of SCH. The cluster head estimates the residual energy of all cluster members and chooses the high-energy node as CH. To find nodes at a maximum threshold distance of D_{th}, CH is used to enumerate the nodes in the cluster. Nodes are considered spatially analogous if the distance between them (D) < D_{th}. The similarity between nearby nodes and SCH data is established by recording similarities in two stages. The first stage involves determining the similarity in magnitude and trend. Let *x* represent a time series for node x (x_1, x_2,..., x_n); let *y* represent a time series for node *y* (y_1, y_2,..., y_n). Using the Euclidean method the distance between the two

time series, expressed as *d(x,y)*, is used to indicate the size similarity between the two time series:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad \text{.....(3)}$$

The Pearson's correlation coefficient can be employed to determine the existing linear correlation between the given time series, which is calculated using the following formula.:

$$r = \frac{\Sigma(x-\underline{x})(y-\underline{y})}{\sqrt{\Sigma(x-\underline{x})^2 \Sigma(y-\underline{y})^2}} \text{.....(4)}$$

Here an α-like data series is shown for $T(x,y) > 0.9$ and $d(x,y) < \alpha$. When both criteria are satisfied, node y will be included in the subcluster headed by node x, which is identified by CH. Subclusters are formed such that members within subcluster α are members of subcluster α. They are similar, with one representative per subcluster. All representative data have the same α, but since CH only communicates with representative data, the error bound for the remaining data can be approximated by α, reducing overall sensor power consumption.

Leveraging Temporal Correlation

Continuous similarity of nodes depends on their temporal correlation at a point in time. Over time, this temporal correlation is burdened by a significant amount of redundant data. Temporal correlations between data are evaluated within a defined precision to identify subsets of sensor readings and reduce energy consumption. Base stations avoid predicting data that has already been delivered to reduce communication. The analysis of sensor data would be performed in the temporal domain of digital filter, making it simpler to anticipate future values. LMS filters can be used to draw conclusions about the short-term linearity of a signal. Based on previous inferences, recent data history can be used to predict linear combinations of future data.

Prediction-based reporting can be used to reduce data within clustered data attack structures. Identical prediction filters are defined in sensor nodes and CH. If the deviation between the actual value detected by the sensor node and the predicted filtered value at a specific time *t* is below a certain threshold, no data will be transmitted. However, if the difference exceeds the threshold, the data will be sent to the CH. Thus, only a fraction of the data is broadcasted.

Forecast-based reportage is performed in three modes. At every sample time *t*, the sensor node sends data to the cluster head. The prediction engine promptly revises its coefficients to convergence depending on the deviations. This is referred to as the commencement mode. When the error limit for β exceeds the error limit for deviation, the predictions of M-continuous predictions converge, and the filter transitions to the detached mode, where the filter model connects to the cluster head. The Temporal Correlation based on *Turan's* theorem to overcome large subset problems:

$$C_b \leq \alpha \geq C_p$$

Span equivalence can be computed using *Turan's* graphs, as described by *Reibiger*[10]:

$$\frac{k-1}{2k}(2 \propto -1)p^2$$

At each sample time *t*, the actual found value in standalone mode is compared to the filter's predicted value. The filter model is deemed precise at time *t* if the deviation from the actual threshold is small. No data transfer takes place in this case. On the other hand, the CH model is widely accepted, and predictors can be computed and viewed as a time approximation of the actual observations. No communication between sensor nodes and CH is required if the observations of the model are predicted accurately. Standard mode is switched to once the error surpasses β. And transfers data to CH in normal mode. The prediction filter weights are adjusted to converge the predictions to the desired value. The filter model receives CH updates and returns to standalone mode when the predictions converge.

Leveraging Spatio-Temporal Correlation

In a traditional prediction-based reporting method, the CH receives data from all CH members using independent prediction filters. A closely spaced and highly correlated data source is characterized by a subcluster of this proposed work and a single SCH node. Thus, the prediction filter's predicted sub-cluster nodes are deemed adequate for the whole sub-cluster. Here sub-cluster head and cluster head comparisons are analysed by prediction-based reports using common models. Due to the close correlation between neighbouring representative nodes, the

10

data produced by the representative nodes are similar in size and trend to the entire subcluster. A collection of representatives characterizes the data fit across subclusters, whose error bounds fall within α. Non-sampler nodes have no error assurance. We predicted an innovative population double-prediction methodology that aids to detect spatial anomalies during data acquisition and corrects them using a model-driven push scheme. For subclusters, predictive models are built by each node based on observed data. Past observations use correlations to predict expected values seen at subsequent times *t.* The model and its parameters are received by the CH and other subcluster members.

At each sample time *t*, the values projected by the model are assessed against the actual measured values. If the threshold is exceeded by the difference between the actual collected data and the model projected value, then the model is said to be α-like for that current period. In this case, none of the subcluster members send data. Since CH knows the model, it can predict the value and use it as an alpha approximation of the actual observed value for a given sensor node. If model predictions and observations are similar, no communication between subcluster members and CH is required. Conversely, when the variance between the model projected data and the collected data exceeds the set limit, then the collected value is transmitted to the cluster head. In this way, by detecting trend deviations, sub-cluster members only send data when there is a deviation in the values of the common model's predicted values. The proposed system lowers communication operating cost by having only one sub-cluster head per sub-cluster and eliminates the possibility of missing deviant patterns in the other nodes' data. The system uses VSS-nLMS-based prediction filters to build the forecasting model; it offers high accuracy and it's computationally efficient. The model's parameters are continuously updated to reflect the current trend, even when trends change. The integration of feedback among CHs, representatives, and other sub-cluster members enhances data reliability and conserves energy.

Conclusion

In summary, WSNs can be utilized in AQI monitoring by providing real-time data transmission, energy efficiency, scalability, and cost-effectiveness. The use of WSNs can improve the accuracy and reliability of AQI data and provide valuable insights into air quality patterns in a given area.

The SAMI system is designed to reduce data variations and improve the accuracy of the collected data through two stages of reduction. It was evaluated based on data reduction and the mean absolute deviation of the data. This leads to a reduction in both spatially and

temporally redundant data, with a data reduction that is many times better than previous systems. SAMI ensures user-specified error thresholds in both spatial and temporal dimensions and provides better system performance and fault tolerance. The system is scalable, and the impact of various parameters was analysed. A mean absolute deviation of 0.07°C resulted in a 75% reduction in data transfer. Future work will concentrate on tuning the spatial and temporal error thresholds in response to changes in data and spatial variability.