Data Compression in Sensor Networks

Amol Deshpande

University of Maryland, `http://www.cs.umd.edu/~amol`

SYNONYMS

Distributed source coding; Correlated data collection; Data suppression

DEFINITION

Data compression issues arise in a sensor network when designing protocols for efficiently collecting all data observed by the sensor nodes at an Internet-connected base station. More formally, let X_i denote an attribute being observed by a node in the sensor network – X_i may be an environmental property being sensed by the node (e.g., *temperature*), or it may be the result of an operation on the sensed values (e.g., in an anomaly-detection application, the sensor node may continuously evaluate a filter such as "$temperature > 100$" on the observed values). The goal is to design an energy-efficient protocol to periodically collect the observed values of all such attributes (denoted $X_1, \cdots, X_n$) at the base station, at a frequency specified by the user. In many cases, a bounded-error approximation might be acceptable, ie., the reported values may only be required to be within $\pm\epsilon$ of the observed values, for a given ϵ. The typical optimization metric is the total energy expended during the data collection process, commonly approximated by the total communication cost; however, metrics such as minimizing the maximum energy consumption across all nodes or maximizing the lifetime of the sensor network may also be appropriate in some settings.

MAIN TEXT

The key issue in designing data collection protocols is modeling and exploiting the strong spatio-temporal correlations present in most sensor networks. Let X_i^t be a random variable that denotes the value of X_i at time t (assuming time is discrete), and let $H(X_i^t)$ denote the *information entropy* of X_i^t. In most sensor network deployments, especially in environmental monitoring applications, the data generated by the sensor nodes is typically highly correlated both in time and in space — in other words, $H(X_i^{t+1}|X_i^t) \ll H(X_i^{t+1})$, and $H(X_1^t, \cdots, X_n^t) \ll H(X_1^t) + \cdots H(X_n^t)$. These correlations can usually be captured quite easily by constructing predictive models using either prior domain knowledge or historical data traces. However, because of the distributed nature of data generation in sensor networks, and the resource-constrained nature of sensor nodes, traditional data compression techniques cannot be easily adapted to exploit such correlations.

The distributed nature of data generation has been well-studied in the literature under the name of *Distributed Source Coding*, whose foundations were laid almost 30 years ago by Slepian and Wolf [6]. Their seminal work proves that it is theoretically possible to encode the correlated information generated by distributed data sources at the rate of their joint entropy even if *the data sources do not communicate with each other*. However this result is non-constructive, and constructive techniques are known only for a few specific distributions [4]. More importantly, these techniques require precise and perfect knowledge of the correlations. This may not be acceptable in practical sensor networks, where deviations from the modeled correlations must be captured accurately. Pattem et al. [3] and Chu et al. [2], among others, propose practical data collection protocols that exploit the spatio-temporal correlations while guaranteeing correctness; however, these protocols may exploit only some of the correlations, and further require the sensor nodes to communicate with each other (thus increasing the overall cost).

In many cases, it may not be feasible to construct a predictive model over the sensor network attributes, as required by the above approach, because of mobility, high failure rates or inherently unpredictable nature of the monitored phenomena. Suppression-based protocols, that monitor local constraints and report to the base station only when the constraints are violated, may be used instead in such scenarios [5].

Sensor networks, especially wireless sensor networks, exhibit other significant peculiarities that make the

data collection problem challenging. First, sensor nodes are typically computationally constrained and have limited memories. As a result, it may not be feasible to run sophisticated data compression algorithms on them.

Second, the communication in wireless sensor networks is typically done in a broadcast manner – when a node transmits a message, all nodes within the radio range can receive the message. This enables many optimizations that would not be possible in an one-to-one communication model.

Third, sensor networks typically exhibit an extreme asymmetry in the computation and communication capabilities of the sensor nodes compared to the base station. This motivates the design of pull-based data collection techniques where the base station takes an active role in the process. Adler [1] proposes such a technique for a one-hop sensor network. The proposed algorithm achieves the information-theoretical lower bound on the number of bits *sent* by the sensor nodes, while at the same time offloading most of the compute-intensive work to the base station. However, the number of bits *received* by the sensor nodes may be very high.

Finally, sensor networks typically exhibit high message loss and sensor failure rates. Designing robust and fault-tolerant protocols with provable guarantees is a challenge in such an environment.

CROSS REFERENCE

Continuous queries in sensor networks; in-network query processing; model-based querying in sensor networks; data aggregation in sensor networks; data fusion in sensor networks.

REFERENCES

[1] M. Adler. Collecting correlated information from a sensor network. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2005.

[2] D. Chu, A. Deshpande, J. Hellerstein and W. Hong. Approximate data collection in sensor networks using probabilistic models. In *Proceedings of the International Conference on Data Engineering (ICDE)*, 2006.

[3] S. Pattem, B. Krishnamachari, and R. Govindan. The impact of spatial correlation on routing with compression in wireless sensor networks. In *Proceedings of the International Conference on Information Processing in Sensor Networks (IPSN)*, 2004.

[4] S. Pradhan and K. Ramchandran. Distributed source coding using syndromes (DISCUS): Design and construction. *IEEE Transactions on Information Theory*, 49(3), 2003.

[5] A. Silberstein, G. Puggioni, A. Gelfand, K. Munagala and J. Yang. Making Sense of Suppressions and Failures in Sensor Data: A Bayesian Approach. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, 2007.

[6] D. Slepian and J. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4), 1973.