A Hierarchical Representation for Recording Semantically Condensed Data from Physically Massive Data Out of Sensor Networks Geographically Dispersed

MinHwan Ok

Korea Railroad Research Institute,
Woulam, Uiwang, Gyeonggi, Korea
panflute@informatics.krri.re.kr

Abstract. A number of sensor networks may produce a huge amount of data, and there has been a necessity the data are processed in a single system. However the data could early overwhelm the database of the system. This work introduces a condensing method to reduce the amount of data exploiting its semantics. The condensing reduces the amount of data to be transmitted and stored, by condensing the data according to semantics shared among servers. The briefed data could diminish the load of applications running on resource-constrained devices in pervasive computing.

Keywords: Sensor Network, Distributed Databases Application, Semantic Condensing.

1 Introduction

Many attributes of the physical phenomenon surrounding people such as air temperature, humidity, and dust density in public facilities are becoming online by sensor networks in pervasive computing paradigm. Since those sensors are geographically dispersed and produce data at predefined rates, the sensor networks would require a distributed data management in a regional or nationwide scale. In these sensor networks, the sensor data is stored near its source, and data processing and filtering are pushed to the edges. Similarly, on the supposition the sensor nodes are tiny hardware suffering energy shortage or etc., queries for the data captured by sensor nodes are preferably processed the sink nodes. Such architecture reduces bandwidth requirements and enables parallel processing of sensor feeds[1].

While many distributed systems are geared toward workloads that are read-intensive, low volume, or not time critical, the distributed systems with these sensor networks will be write-intensive, high volume, and often time critical. Since the volume of sensor data becomes enormous if they are congregated nationwide, those data do not seem accommodated in a few database systems, in the form of raw data. In this work, a condensing method is proposed to reduce the amount of data exploiting its semantics. The condensing reduces the amount of data to be transmitted and stored, by condensing the data according to semantics shared among servers. The

R. Meersman, P. Herrero, and T. Dillon (Eds.): OTM 2009 Workshops, LNCS 5872, pp. 69–76, 2009.

building and updating processes are suggested for hierarchically distributed sensor databases, exploiting the merit of semantic condensing. The underlying database is a specific database including TinyDB, and COUGAR[2], which are the common database management systems for sensor networks. Distributed-stream-processing engines such as Aurora, Medusa, Borealis, Telegraph-CQ, and HiFi are the future candidates.

An early system designed to provide a worldwide sensor database system is IrisNet[3]. It supports distributed XML processing over a worldwide collection of multimedia sensor nodes, and addresses a number of fault-tolerance and resource-sharing issues[4]. However there has not been any related work based on a concept similar to a condensing method to brief the original data, introduced in this work, in our best knowledge.

2 Condensing Ranges into Semantic Values from Linear Data

The sensor captures the attribute states of a certain physical phenomenon by time, and nowadays, many applications use the sensors that produce a series of one variable such as temperature, humidity, density or pressure. The produced data are values captured according to time and this type of data is called *linear data* in this work, as continuous values constitute the data of one variable.

Due to large amount of data, including that of energy consumption, etc., most sensors capture values between intervals, for specific durations, or at different rate along time. Although the capture time may not be continuous, the produced data is a linear data and it is stored in a database attached to the sensor network. Suppose there are several sensor networks an organization operates, and a number of the organizations are located in a region. Consider a sort of attribute of a certain physical phenomenon should be attended in a region, an air temperature higher than the 50 degrees centigrade in the rooms of the buildings for example, so the regional fire brigades could be notified and turn out. In this case the data produced in the region should be huge and if a database of the server covered the region, it would be early overwhelmed. If the regional server does not have its database but organizations' databases only store the data, the regional server need query to every database of each organization or every organizational server should prepare their data to the regional server periodically, for the regional command of the fire brigades.

It is impractical to have the region database in capacity equal to summated capacities of all the organization databases and thus the amount of the aggregate data should be reduced. In reducing the amount, some information may be lost but in a way maintaining the useful ones. Fig. 1 depicts the concept of *condensing*, reducing the amount, from the sink node of bottom sensor nodes to the topmost central server with the coverage of a nationwide.

There could be three modes in capturing values at sensor nodes. The first mode is continuous capturing, which the sensor captures values continuously. The second mode is discrete capturing, which the sensor captures values at specific times,

i.e. periodically. The third mode is the composite capturing, which the sensor captures values continuously for durations with time gaps between the durations. Continuous capturing causes early energy drain and the other mode is preferred in many cases. The sensors produce linear data in each mode, and their data are collected in a sink node of the sensor network. The sink nodes send their data to the organizational server that has the organization database of the sensor networks the organization operates. Fig.2 shows this procedure that a region database is updated every interval.

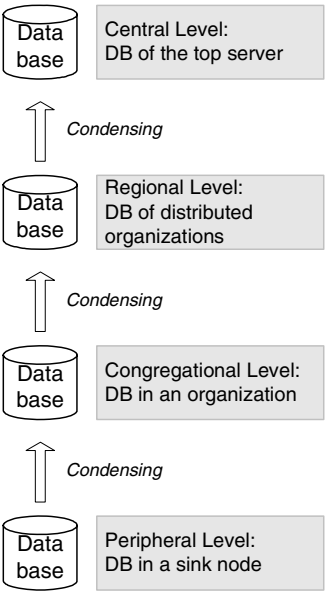

Fig. 1. Hierarchically upward condensing

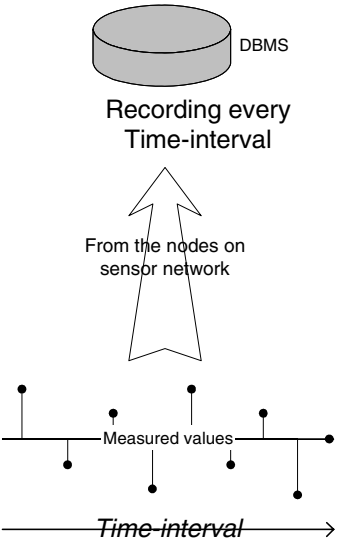

Fig. 2. Linear data from sensor node

In reality, all of the large amounts of data are not necessary for every query. Thus the top server may not have stored all the data. A sensor data is a series of values in domain the sensor is able to capture. A number of physical phenomena have a characteristic that the captured values has their semantics defined according to the degree they belong to. For the regional command of fire brigades, for example, it is normal temperature below 50 degrees centigrade, but abnormal temperature over 50 degrees centigrade. Therefore whether the air temperature is over or below 50 could be the only interest of the regional command and then the region database may merely stores NORMAL or ABNORMAL. Near the physical phenomenon, at the organization of the sensor network, more distinctions should be necessary to immediately cope with imminent fire, and the organization database might store 4 statuses, ORDINARY, NOTICE, WARNING, DANGER, which is less condensed data. Each status has a meaning of the contiguous range the captured value could

belong to. Further those statuses are not necessarily stored continually since the start time of the status and the status value compose a tuple as a sufficient data. In general,

$$T_i = \{t_S, S\}, \tag{1}$$

where T_i is the tuple of the sensor node with ID *i*, S is the status value, and t_S is the start time of the status. This is named as *Semantic Condensing* by ranges of captured values. If the status values are binary, even the status value could be omitted from the tuple, resulting a more condensed form. The database of Central Level in Fig.2 could store such a special tuple in the assumption the status must be NORMAL in the beginning.

3 Hierarchically Distributed Sensor Databases

The distributed databases of the organizations are the essential entities in comprising an entire sensor database system. According to the coverage of interest, i.e., nationwide, regional, or organizational, the client may make request on the aggregate data to the central server, a regional server, or an organizational server, respectively. In this system, either sink nodes or the organizational server should maintain its database with uncondensed data so it can reply to queries for more detailed data delivered from the higher level. In building the hierarchically distributed databases depicted in Fig. 4, the cooperation between databases of the higher-level and the lower level is as follows;

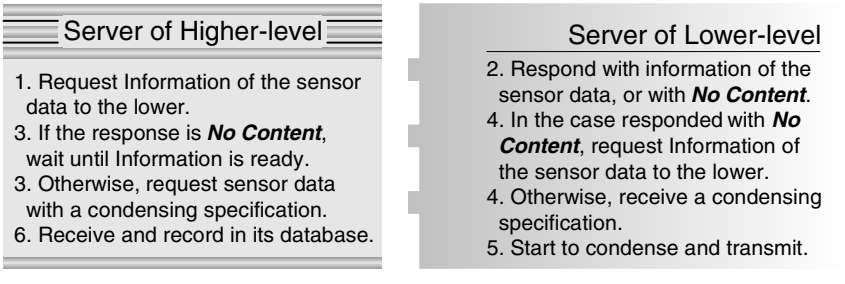

Fig. 3. Building process of hierarchically distributed sensor databases

The information the server of the higher-level requests are an upper limit, a lower limit and the resolution of the variable. The requested data is sent after condensing at the server of the lower level with other upper limit, lower limit and resolution as specified from the server of the higher-level. The building process is initiated from the central level to the organizational or peripheral level, and completed from the lower-levels to the higher-levels, to build the hierarchically distributed sensor databases. Multiple hierarchies of databases could be built with other sets of condensing specifications, as shown in Fig. 4.

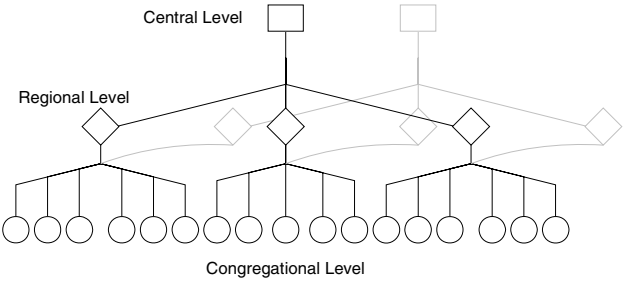

Fig. 4. Building process is completed from the lower-levels

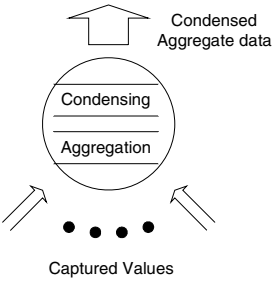

Fig. 5. Condensing captured values

Once a database of the higher-level is built out of the databases of the lower-level, updates for new data should follow. Basically the time to send the new data follows the expression (1) to the database of the higher-level, i.e., when the status has changed. In applications the time the status changes is not crucial, the new data could be sent every a certain time-interval. Since the data being condensed are linear data based on time, the time gap between the latest records increases between the levels as the level is closer to the central level. For this reason, the building process may be initiated more than once if the time of the latest record is far from current time in the database of the central level. Only records of the omitted time are appended during this partial building.

4 Reduced Data by Semantic Condensing along the Hierarchy

For many applications, the range of values is equally significant with the raw data gathered from sensor networks. In the cases, the distinction between aggregated data has the meaning equal to the exact value. For example, what high the temperature is equally meaningful to the exact temperature for the fire brigades. Furthermore, there could be intrinsic errors in the values captured. It is more evident in the cases some actuators are connected to the sensor networks to autonomously react to the data, i.e., whether the temperature is over 50 degrees centigrade.

Conventionally the organization in the higher level uses briefed data of the organizations in the lower level, since it should manage all the events concentrated from the lower level. As higher in the hierarchy, the data need be more condensed in a way not losing its semantic meaning. Either sink nodes or the organizational server should maintain its database with raw data. Consider r levels in the hierarchy of the databases and level r is the topmost. The reducing rate at level 2 is the resolution of condensing from the raw data. The reducing rate at level 3 is the ratio of the resolution of level 2 to the resolution of level 3, and so on. This reducing rate enlarges along the hierarchy, as shown in Table. 1, which is an example the resolution of condensing is requested by 50% at each level.

Table 1. The reducing rate enlarges along the hierarchy

Level of Database	Resolution of Condensing	Reducing Rate
1st-level	100%	1
2nd-level	50%	$1/2$
3rd-level	50%	$1/4$
⋮	⋮	⋮
r th-level	*50%*	$1/2^{r-1}$

In general, the reducing rate of data to be transmitted and stored, S_r, is as follows, at least, in the hierarchy of the distributed databases;

$$S_r \geq \prod_{h=1}^{r} R_h \,, \tag{2}$$

where R_h is the resolution of data at the *h*-th rank in the hierarchy, supposing the bottommost is the 1st rank and the resolution of higher-level is lower than that of lower-level, for every level. While the status value does not change, i.e., the newly captured value is still in the same range, the reducing rate gets much higher.

The reducing rate is acquired at the cost of losing much data not interested. In addition, other parallel hierarchies could be necessitated for other interests. Assume a new application is required for other purpose to process data from the same sensor networks, as shown in Fig. 4. The new application requests different resolutions of condensing in the middle of the hierarchy, thus builds a new hierarchy different from the middle. The data of the new hierarchy of other interests also take place in the databases. In this scenario, the effect of semantic condensing decreases, however it should be efficient to condense the data semantically than not to condense, unless the amount of condensed data is larger than that of raw data by the number of hierarchies built excessively. Semantic condensing has another merit of indexing to the data, similar to Domain Name Service, which means faster access to the data.

5 Related Works

The nationwide sensor database system of this work is similar with the concept of *Virtual Sensor*[5]. Virtual sensors abstract from implementation details of access to sensor data and define the data stream processing to be performed. Local and remote virtual sensors, their data streams and the associated query processing can be combined in arbitrary ways and thus enable the user to build a data-oriented 'Sensor Internet' consisting of sensor networks connected via a global sensor networks. The coalition of virtual sensors is *Virtual Sensor Networks* (VSN)[6] to provide protocol support for the formation, usage, adaptation and maintenance of subsets of sensors collaborating on specific tasks. Its example introduced the functionality including the support for nodes to join and leave VSNs, broadcast within a VSN, and merging of VSNs.

While those works have proposed the concept, mechanisms, and benefits of using VSN, an XML extension technique called Stream Feed[7] has addressed the sensor data-stream and evaluated their technique against the large streaming data object. As the sampling interval decreases the number of clients reduced, and as the network is deeper the latency increased. They are natural results but a big obstacle in creating an application of sensor database with a nationwide coverage.

6 Summary with Future Work

The hierarchically distributed databases store the condensed data of one kind of sensors. Some of servers could request exchanging their condensed data each other to form new data by combining the data of different kinds of sensors. The organization could operate the networks of other kind of sensors, i.e. the sensor for temperature and one for smoke. Captured data of smoke sensors should be helpful in detecting the occasion of fire, in the previously described example. The reduced amount of data lessens transmissions over network, and should be also helpful in exchanging the condensed data.

A condensing method is proposed to reduce the amount of data to be transmitted and stored, by condensing the data according to semantics. The building and updating processes are suggested for hierarchically distributed sensor databases. Although only the method of condensing ranges into semantic values is addressed in this work, there could be another method of semantic condensing by purposes. If there is a set of specific thresholds on one variable and a semantic value is assigned when the value captured is greater than one of the threshold, the status value should be one of the semantic value. In this case the set of specific thresholds are the information for semantic condensing, not an upper limit, a lower limit nor the resolution. Semantic values resulted from a complex combination of conditions also possible, and thus this is why the semantic condensing is different from a generic coding.

The reduced size of data becomes an advantage in pervasive computing. The briefed data could diminish the load of applications running on resource-constrained devices, such as handheld devices, by semantic condensing. It is also preferable in creating applications of nationwide smart space for use in pervasive computing.

References

1. Balazinska, M., Deshpande, A., Franklin, M.J., Gibbons, P., Gary, J., Nath, S., Hansen, M., Leibhold, M., Szalay, A., Tao, V.: Data Management in the Worldwide Sensor Web. IEEE Perv. Comp. 6(2), 10–20 (2007)
2. Henricksen, K., Robinson, R.: A Survey of Middleware for Sensor Networks: State-of-the-Art and Future Directions. In: International workshop on Middleware for sensor networks, pp. 60–65. ACM, New York (2006)
3. Campbell, J., Gibbons, P.B., Nath, S.: IrisNet: An Internet-Scale Architecture for Multimedia Sensors. In: Annual ACM international conference on Multimedia, pp. 81–88. ACM, New York (2005)

4. Deshpande, A., Nath, S., Gibbons, P.B., Seshan, S.: Cache-and-query for wide area sensor databases. In: ACM SIGMOD international conference, pp. 503–514. ACM, New York (2003)
5. Aberer, K., Hauswirth, M., Salehi, A.: Infrastructure for Data Processing in Large-Scale Interconnected Sensor Networks. In: International Conference on Mobile Data Management, pp. 198–205. IEEE, Mannheim (2007)
6. Jayasumana, A.P., Han, Q.: Virtual Sensor Networks - A Resource Efficient Approach for Concurrent Applications. In: International Conference on Information Technology, pp. 111–115. IEEE CS, Las Vegas (2007)
7. Dickerson, R., Lu, J., Lu, J., Whitehouse, K.: Stream Feeds - An Abstraction for the World Wide Sensor Web. In: Floerkemeier, C., Langheinrich, M., Fleisch, E., Mattern, F., Sarma, S.E. (eds.) IOT 2008. LNCS, vol. 4952, pp. 360–375. Springer, Heidelberg (2008)