

A methodology to optimize query in wireless sensor networks using historical data

Haroon Malik · Ahsan Samad Malik ·
Chanchal K. Roy

Received: 10 January 2011 / Accepted: 13 May 2011 / Published online: 11 June 2011
© Springer-Verlag 2011

Abstract In wireless sensor networks (WSN), a query is commonly used for collecting periodical data from the objects under monitoring. Amount of sensory data drawn across WSNs by a query can significantly impact WSN's power consumption and its lifetime, since WSNs are battery operated. We present a novel methodology to construct an optimal query containing fewer sensory attributes as compared to a standard query, thereby reducing the sensory traffic in WSN. Our methodology employs a statistical technique, principal component analysis on historical traces of sensory data to automatically identify important attributes among the correlated ones. The optimal query containing reduced set of sensory attributes, guarantees at least 25% reduction in energy consumption of WSN with respect to a standard query. Furthermore, from reduced set of data reported by optimal query, the methodology synthesizes complete set of sensory data at a base station (reporting unit with surplus power supply). We validated the effectiveness of our methodology with real world sensor data. The result shows that our methodology can synthesize complete set of sensory data analogues to standard query with 93% accuracy.

Keywords Wireless sensor network · Data reduction · Principal component analysis

1 Introduction

A wireless sensor network (WSN) comprises of numerous sensor devices, commonly known as motes,¹ which usually consists of several sensors that are used for monitoring the physical entities such as temperature, light, motion, metallic objects, and humidity (Tik et al. 2009; Bajaber and Awan 2010). A WSN consists of hundreds or thousands of sensor motes that have the ability to communicate either with each other or directly with the base-station (BS). A base-station is a fixed node or a mobile node that is capable of connecting the sensor network to an existing communications infrastructure or to the Internet. The BS is equipped with a surplus power supply and can carry out computational-intensive operations. On the other hand, motes are equipped with a limited power supply (batteries). The motes can be deployed at hard to reach areas such as glaciers, deserts and battle fields; thus it is hard to replenish their batteries. Hence, energy efficient communication mechanisms are required to reduce the energy consumption and prolong the lifetime of WSNs. Transmission of data is the most expensive activity for a sensor mote. According to (Naoto and Shahram 2005), transmitting one bit via radio medium is at least as 480 times power consuming compare to performing one addition instruction by a mote. Moreover, the energy consumed in transmitting over large distance is proportional to distance (Heinzelman et al. 2000). Thus, in order to avoid direct transmission to BS, many

H. Malik (✉)
School of Computing, Queen's University, Kingston, Canada
e-mail: malik@cs.queensu.ca

A. S. Malik
Ghulam Ishaq Khan Institute of Engineering and Technology,
Topi, Pakistan
e-mail: u2003015@giki.edu.pk

C. K. Roy
Department of Computer Science, University of Saskatchewan,
Saskatoon, Canada
e-mail: chanchal.roy@usask.ca

¹ The term mote was coined by researchers involved in 'smartdust' project at Berkeley.

energy efficient routing algorithms based on multi-hop routing paradigm have been proposed (Uchida et al. 2010; Hiyama et al. 2010; Nose et al. 2010). Furthermore, for reducing the in-network sensory traffic, several data reduction techniques such as packet merging, data aggregation and compression have been proposed for WSNs.

1.1 WSN as distributed database

At a higher level, WSNs can be modeled as a distributed database where every mote is a database. The attributes of the database are the types of sensors mounted on the mote. A user can query the database (mote) to retrieve the values of the attributes of interest, such as pressure, humidity, temperature, longitude, latitude, etc. In a traditional database system, queries are used to search for data contained in a persistent storage repository. In a WSN, a database consists of the environmental data that can be measured/acquired by the transducers available on the sensor motes. Queries instruct sensor motes on the management, filtering, and processing of the data acquired from the environment. The software running on the motes allows data to be acquired when needed from the environment, exactly in the way that traditional database software allows data to be accessed from disks. In a wireless sensor network, data are not stored within a device such as disk; environmental data are acquired by transducers of nodes when needed, in accordance with the query that the WSN process. New data are available every time a transducer is activated which needs to be reported back to the command center or BS for analysis. The data consists of list of sensors and the values of their related attributes. Once data are reported to the BS, it is stored either in WSN specific databases systems such as Tiny DB and COUGAR or as log traces.

The data collected by the sensors is usually highly redundant. Thus, a well designed query, not reporting the irrelevant, redundant and highly correlated sensory data back to BS can further increase the impact of data reduction (DR) and routing techniques. This paper presents a methodology to optimize the construction of such optimal queries.

The key novelty of the methodology lies in exploiting strong correlation among the observed historical values (stored in sensor database) for a query, in order to reduce the sensing and communication overhead. Thus by moving smaller subset of un-correlated but representative samples along WSN, i.e., temporal and spatio-temporal correlation concerning one or more sensor attributes (e.g., correlations between humidity and temperature reading of a nearby nodes over a time period) substantial energy saving is achieved. The representative samples can then be processed to reproduce the data of the correlated attributes at

centralized more power unit, i.e., BS by using association rules (Bronnimann et al. 2003).

The contributions of the paper are as follows:

C1 We apply statistical methods to reduce the dimensionality of the observed sensory data.

C2 We automate the ranking and identify the attributes according to their importance for the query.

C3 We empirically validate our proposed approach through a case study with a real sensor network.

The remainder of the paper is organized as follows. Section 2 introduces the query for wireless sensor networks. Section 3 briefly describes related work. In Sect. 4 we present our methodology followed by a case study in Sect. 5. The performance of our approach is evaluated in Sect. 6. Finally, we conclude the paper by outlining few future works in Sect. 7.

2 Query in sensor network

SQL-based query language, which consists of a SELECT-FROM-WHERE clause, is commonly accepted and widely used in specifying queries for a sensor network. However, sensor network has its own characteristics, and therefore extensions must be made to the basic SQL query. We provide an example WSN query in the form of extended SQL for this instructive section:

The example query as shown in Table 1 for WSN specifies that after 5 min from now (the time at which query is diffused into WSN), the motes in Area_C should measure the temperature, pressure, humidity, photonic attributes per second for 120 times, and the resulting temperature should be reported within 10 min after the last sensor sampling is completed. Delay in WHILE clause is as the QoS constraint, and the SAMPLE clause specifies the temporal information. From delay and sample time, we can easily get the latest report time. With respect to various temporal features, a query can be either one-time query or continuous query. The one-time query is to sense the object only once at a specific moment of time, and is used to take a snapshot of the phenomena under observation. In one-time query, there are no INTERVAL and LOOP fields. This paper concentrates on the optimization of periodical query during the fourth (last) stage of query execution, i.e., result collection. In particular, we mainly consider the situation where there is a delay constraint and no operations (aggregation, fusion) are to be performed on the collected data for the query. In such case, the query result is exactly a stream of sampling data flowing from the network to the query point, i.e., the sink. This sort of query is very popular in real world applications like environment monitoring and healthcare (Madden and Hellerstein 2002).

Table 1 A sample WSN query

<i>SELECT</i>	Temperature, pressure, photonic, humidity
<i>FROM</i>	Sensor as s
<i>WHERE</i>	s.location = Area_C
<i>WHILE</i>	Delay <10 min
<i>SAMPLE</i>	
<i>ON</i>	Now +5 min
<i>INTERVAL</i>	1 s
<i>LOOP</i>	120

3 Related work

We discuss two areas of related work.

3.1 Query processing

There has been a lot of work on query processing in the area of distributed database systems (Yu and Chang 2000; Ozsy and Valduriez 1991; Raghunathan et al. 2002). However, as discussed in Sect. 1, there are few major differences between sensor networks and traditional distributed database systems. The traditional distributed systems do not consider the physical limitation of sensor networks (Shatdal and Naughton 1995). Most related work for WSN query processing deals with distributed query processing with an exception of Cougar.

COUGAR (Bonnet et al. 2001) proposed an enhanced centralized architecture where the system extracts from network the only data required by user queries instead of all sensed data. TinyDB (Madden and Franklin 2002; Madden and Hellerstein 2002) attempts to reduce power consumption of sensor nodes that are required to route and deliver data to a base station by taking part of processing from base station into the network. TinyDB uses a cost-based optimizer to choose a query plan which has the estimated lowest overall power consumption. Each node in TinyDB consists of a catalog of metadata that describes its local attributes, events, user-defined functions as well as the cost of processing and delivering data. These metadata are periodically copied to the root of the network for use by the query optimizer. Although TinyDB has taken power consumption for one of its factors of selecting the most suitable execution plan, it did not consider how the execution plan will change when network topology changes.

The DIM (DIM Project 2007) proposed to build an in-network distributed data structure for efficiently answering multidimensional range queries. The main challenge of this project was to manage the high volume of sensor output, and present information in a high-level form useful for achieving decision support and planning.

ICEDB (Zhang et al. 2007) is a buffering mechanism for managing the delivery of query results from mobile nodes to a base station and the queries from base station to mobile nodes. For this mechanism, the local query processor continues to gather, store and process the collected data even during the periods with poor connectivity. After the connectivity resumes, those collected data are sent in the order of perceived importance.

Another dimension of query processing in WSN, which also stems out from traditional database practice is evaluating the top- k queries. The top- k queries request the list of k sensor nodes with the highest or lowest readings. There exist many situations, where one is interested in monitoring extreme and atypical behavior. For example, finding the highest pressure points in a pipeline, the largest temperature values in a building or patch of forest, the points of most intense vibration in a bridge, etc. Getting the required sensory data only from the nodes that have extreme values (largest, highest and intense) can greatly reduce the energy consumption of the WSN.

FILA (Wu et al. 2007) is an energy efficient top- k monitoring approach. The basic idea of FILA is to install a filter at each sensor node to suppress unnecessary sensor updates. A model-driven data acquisition approach, which suggests using models such as multivariate gaussians to predict sensor readings, is proposed (Deshpande et al. 2004). These models avoid visiting nodes whose readings can be accurately predicted or are unlikely to contribute to the final result. This approach can dramatically reduce the energy consumed by the network, but of course makes results approximate. Instead of using models explicitly, Silberstein et al. (2006) propose to use samples of past sensor readings. The samples are computationally efficient to use in query optimization. It demonstrates the power and flexibility of sampling-based approach by developing a series of top- k query planning algorithms with linear programming.

The Threshold Join Algorithm (TJA) (Yazti et al. 2005) is an efficient top- k query processing algorithm for distributed sensor networks. The TJA uses a non-uniform threshold on the queried attribute in order to minimize the number of tuples that have to be transferred towards the querying node. One of the shortcomings of TJA is that the sensor nodes consume extreme energy to probe sensor reading or update filters. Especially, it becomes worse, when the variation ratio of top- k result is higher (Yeo et al. 2009). Finally, NER (Liu et al. 2010), a query processing method was introduced to resolve the problem of vast queries that do not indicate the query region in multi-user scenario by introducing the event names. The event reporting mechanism employed in NER reports the on-going events to the BS, indicating what and where the event are in the network.

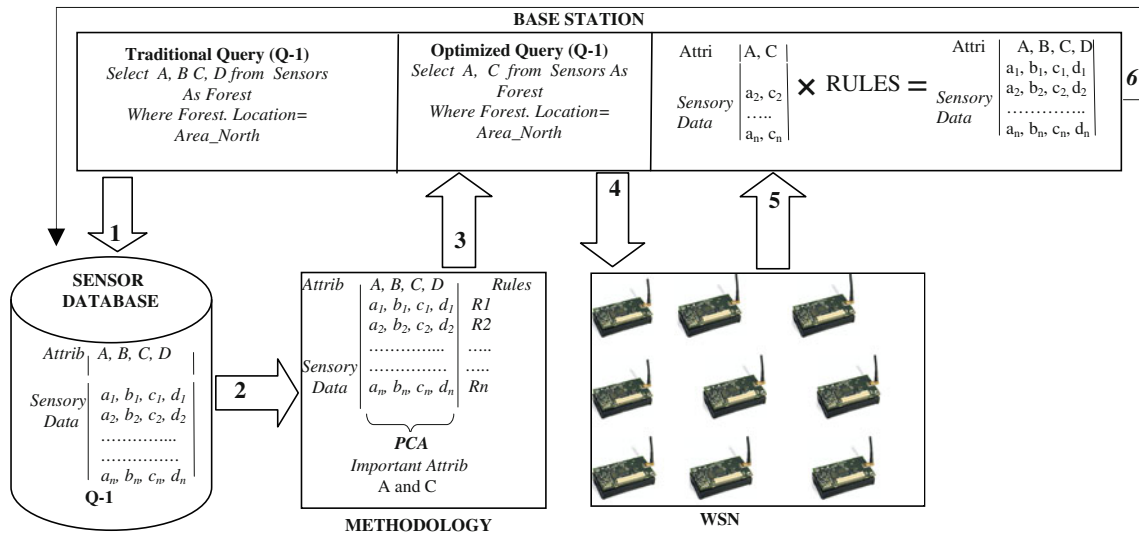


Fig. 1 Workflow of our methodology

3.2 Data reduction techniques

Data reduction is to decrease the size of data that is exchanged between the sensor motes. The idea is straight forward; small amount of data consume less amount of power in transmission. Various data reduction techniques exist in this context, including packet merging, packet compression, data aggregation and data fusion.

Packet merging is a simple data reduction technique, which combines multiple small packets into a big one, without considering the correlations between and the semantics within individual packets (Ju and Cui 2005). Packet compression is to integrate one or multiple packets into reduced packets, by employing suitable data compression algorithms. A number of compression algorithms have been studied for sensor networks (Xu et al. 2003). Work presented by Ju and Cui (2005) shows that total power consumed by sensor a mote is reduced by employing compression algorithm. However, applying data compression can even increase total power consumption based on the type of application data (Xu et al. 2003). This is due to accessing memory during compression execution time. Accessing memory is expensive in terms of energy consumption. A rate–energy–accuracy (R–E–A) tradeoff frame work is proposed for selecting suitable compression algorithms (Chen and Fowler 2004). To the best of our knowledge, constructing optimal query to reduce the sensor data flow through sensor network as part of data reduction has yet not been exercised.

4 The methodology

In this instructive section, we discuss our methodology in action, using a real world sensor trace of 17 sensor

attributes over a period of 2 h for 18 motes. Different WSN applications usually have different requirements for accuracy, energy consumption and information delay. For example, the WSN deployed for a battle field and for a rescue purpose may have short life span (i.e., high energy consumption) but require higher accuracy and real time information. Whereas WSNs deployed for habitat and environmental monitoring as compared to WSNs for battle field and for rescue purposes are typically on large scale consisting of hundreds of motes. In such large WSNs, power efficient execution of long-running queries is the main concern. These networks do not require precise results with high degree of accuracy at the cost of intense computation and short life span. Moreover, the information obtained from such WSNs, i.e., deployed for habitat and environmental monitoring can bear some delay tolerance, i.e., some extra processing required at BS to synthesize information. Our methodology aims for such energy conscious and delay tolerant sensor network.

The workflow of our methodology consists of five simple steps as shown in Fig. 1. We explain these five steps using an example. *Step 1* The result of traditional query (Q-1) containing four attributes (A, B, C and D) as shown in Fig. 1, received at BS is stored in sensor database for a certain period of time. *Step 2* Methodology employs principal component analysis (PCA) technique on the result of traditional query (Q-1) to identify important attributes such as A and C. The methodology then recommends them to BS as shown in Fig. 1. Furthermore in second-step, (R_n) data rules are constructed for Q-1 results using association rule mining. These rules will be utilized latter in step 5 of our methodology workflow. *Step 3* The BS uses the important attributes recommended in step 2 by our methodology to construct an optimal query. The optimal query

has fewer attributes as compared to the standard query; hence lesser data are drawn from WSN thereby reducing the power consumption of the WSN. *Step 4* The BS diffuses optimal query in WSN. *Step 5* The reduced set of sensory data resulting from optimal query is collected at BS which is equipped with surplus power supply. At BS, these sensory data are further treated with rules extracted from historical data in step 2 to synthesize the sensory data of other attributes used in standard query (Q-1), i.e., B and D.

We now explain the steps of our methodology as shown in Fig. 2:

1. *Data preparation* Sensor trace logs need to be prepared to make it suitable for the statistical techniques employed by our methodology. Figure 10 shows a sample log trace of our mote.

The effectiveness of the suggestions generated by our methodology greatly depends on data preparation. The two steps involved in data preparation are:

(a) *Data sanitization* Sensor trace logs need to be filtered from noise, i.e., missing attribute data or an empty attribute. Attribute is missing when mote fails to record an instance of an attribute. An attribute is empty when sensor on mote fails to initialize or gets malfunction.

(b) *Pre-treatment* It converts the data into a format that is understood by the data reduction technique, i.e., PCA.

PCA is a maximum variance projection method (Jolliffe 2002). This means that PCA identifies those variables that have large data spread (variance), ignoring variables with low variance (Box and Box 1996). Mote's attributes have different ranges of numerical values; they have different variance. To eliminate PCA bias towards attributes with a larger variance, we standardized the performance attributes via *Unit Variance scaling* (UV scaling). For each attribute, we standardized it by dividing the samples of each sensory attribute by the attribute's standard deviation. Each scaled attribute then has equal (unit) variance.

2. *Data verification* The second step of our methodology verifies if there exists enough association among mote's data to proceed with the data reduction. In

order to apply PCA, the Kaiser–Meyer–Olkin (KMO) measure should be greater than 0.6 (Kaiser 1974). This measure tests the amount of variance within data that can be explained by a given measure. The KMO measure for our mote's data is 0.789, which indicates PCA is appropriate to apply.

3. *Dimension reduction* We consider the elimination of redundant attributes in sensor query as a dimensionality reduction (DR) problem, where each sensor mote corresponds to a dimension. Many different DR techniques exist, for example based on statistics clustering (factor analysis, alpha, un-weighted least-square) or machine learning (Maximum likelihood, Feature selection, cross entropy, etc.). Among statistical techniques clustering algorithms have been widely used and perform reasonably well on datasets of low dimension, with "low" defined as less than fifteen (Beyer et al. 1999). Unfortunately, we expect to have dimensions over 1,000 in WSNs. Towards this end, we used statistical technique, PCA, known to reduce the sheer volume of both sensor data and attributes and is robust and scalable (Jolliffe 2002). What PCA does is to synthesize new variables called "Principal Components" (PCs). Every PC is independent and uncorrelated with other PCs. We used our custom R scripts and the FactoMineR package dedicated for data mining and multivariate analysis to perform the PCA analysis (Le et al. 2008).

4. *Top_k components* Many sensor attributes have little information value for a certain point of time and hamper effective analysis by adding noise, e.g., the value for every sample of acceleration (accel_x, accel_y) will be constant (0) for a stationary object. Hence such PC (containing noise) needs to be discarded from analysis. Unfortunately the methods known today, does not provide any reliable and automated techniques to identify appropriate top_k principal components. We found it more practical to use "% Cumulative Variability" in selecting the number of top_k component. Table 2 shows that 8 PCs account for 90% of cumulative variability. A cumulative variability of 90% is adequate to explain most of data with minimal loss in information (Jolliffe 2002), enabling us to achieve 52% data reduction by selecting first eight PCs from total of 18 as shown in Table 2.

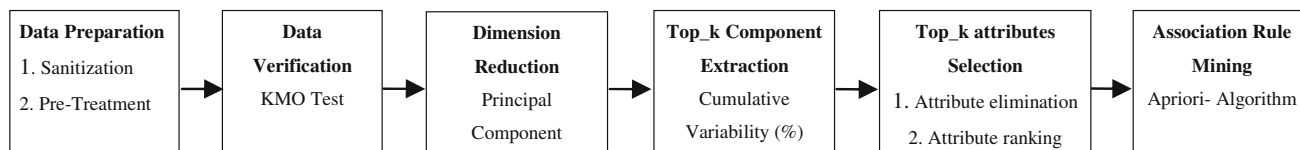


Fig. 2 Steps involved in proposed methodology

Table 2 Principal component analysis

PC	Eigen value	Variability (%)	Cumulative variability (%)
PC1	13.045	44.984	44.984
PC2	3.350	11.552	56.536
PC3	2.839	9.789	66.326
PC4	2.303	7.941	74.267
PC5	1.973	6.803	81.070
PC6	1.306	4.504	85.574
PC7	1.206	4.160	89.734
PC8	0.906	3.125	92.859
⋮	⋮	⋮	⋮
PC18	0.005	0.016	100.00

5. *Top_k attributes* Application engineers are interested in mote's attributes to construct query not principal components. In this step we decompose principal components using eigen vector decomposition technique to map the PCs back to mote's attributes. For each attribute we measure its association to each top_k components. This measure of association is called as "Loadings". Table 3 shows the measure of association (loadings) for few of our mote's attributes. The loading value ranges from $\pm(0-1)$. The attribute "J" and "D" with higher loading values confirm strong association with PC1, whereas attributes like "P" confirm weak association with PC1. In order to remove weakly associated attributes (add no value to the PCs) and to identify and rank the top_k attributes our methodology performs the following two sub steps.

- (a) *Attribute elimination* In this step, the attributes that do not have significant association with their respective top_k dimension are removed. A Norman cut-off criterion is utilized to decide on the level of importance of a variable to corresponding dimension: $\text{Cutoff} = 5.152/[\text{SQRT}(n - 2)]$, where the loading value is considered 5.152 only if we have more than 100 samples and n represents number of samples (Jolliffe 2002).
- (b) *Attribute ranking* In this step, the important top_k attributed belonging to the top_k PC are identified and ranked. Identifying important variables has been made possible in the literature by exploiting loading values in a strict manner. In past literature loading value of 0.7 is used as cutoff criteria to obtain important attributes. Hair et al. call loadings above 0.6 "high" and those below 0.4 "low" to rank important variables, whereas Raubenheimer pointed out 0.4 for the central PC and 0.25 for the other PC (Ozsy and Valduries 1991).

We believe the cut-off level to identify the top_k attributes should not be fixed. It should be tunable on the basis of domain demands. To server this purpose, we incorporated loading as tunable parameter in our second step. Table 4 shows that with loading parameter set to 0.8, our methodology identifies seven out of 18 attributes along different dimensions and with 0.9 only two of the top attributes. Rather than including 18 attributes of interest, including only two attributes "D" and "J" (which are photonic and temperature sensors) in query will solve the purpose. Fetching such small amount of data drastically reduce the energy consumption of mote. Using the sample obtained from these attribute, remaining attributes can be synthesized at BS using association mining which is next step of our methodology. Every attribute (sensor) consumes energy differently. In order to evenly burn energy of the mote the top attribute identified by our methodology for every PC needs to be cycled in a query periodically.

6. *Association rule mining* Identifying top_k attributes along each PC and using them interchangeably in construction of optimal query for energy consumption alone is not enough. There are many corner cases that cannot be justified by simply plugging top_k attribute. What do we mean by corner case? We explain this by our experience during experiments conducted at a botanical garden. From historical traces of sensor data collected from botanical garden, we find that there is strong correlation between the humidity, temperature and photonic (light) sensor data and they all fit into same PC. Using one of these attribute, e.g., temperature to query temperature sensor on mote can greatly save battery power as other two sensors (humidity and photonic) will not be used. At the base station, the data

Table 3 Association between attributes and comp

Attr	Loadings			
	PC1	PC2	PC3	PC4
A	0.399	-0.42	0.055	0.739
J	0.977	0.015	-0.046	-0.134
D	0.979	0.01	-0.043	-0.130
⋮	⋮	⋮	⋮	⋮
P	0.861	-0.122	-0.023	-0.037

Table 4 Top-*k* attributes

Rank	PC	Attributes	Loadings
1	F1	D	0.979
2	F1	J	0.977
3	F1	R	0.866
4	F1	F	0.846
5	F1	P	0.844
6	F2	Q	0.833
7	F2	I	0.812

received for a temperature sensor can be used to synthesize the data of other two sensors (humidity and photonic) by exploiting correlations. On occasions, we have found the strong correlations between these three sensors do not hold true. Once or twice a week, for the purpose of air bath, the air gates are opened for an hour to few days for specific botanical zones in turns. This cause the correlations that exists between the humidity sensor and other two (photonic and temperature) to deviate from the natural correlation that exists for that particular zone in a season. Plugging the humidity attribute in optimal query for that period of time will astray the reproduced results of photonic and temperature sensor (when air gates are opened). We employed Apriori algorithm to construct association rules on the data obtained from historical traces of sensor motes to overcome the corner cases as it has been proved to perform better than other available association rule mining algorithms for very large databases (Rakesh and Ramakrishnan 1994). We used a statistical and data mining tool WEAKA for the purpose. We found 700 historical samples from trace log were sufficient to extract association rules that can synthesize data which is 95% correlated (person-correlation) with actual trace log. For every query, the sensor data reported to base station is compared with the sensory data of the previous query to validate the association rules derived from the historical sensory data. If the rule(s) is violated then the sink instructs the mote to report the data from all the sensor of interest.

5 Case study

To evaluate the effectiveness of our approach, we conducted a case study on the logs obtained from 18 mica motes over the period of 30 days. The Mica2 mote has an 8-bit microcontroller (ATmega 128L) running at 8 MHz, 4 kB of RAM, 128 kB of flash and a 916 MHz radio with 38.4 kB/s maximum theoretical bit rate. The motes were placed in a green house at our university botanical garden exactly one meter apart from each other. All the motes were equipped with a pair of 2,100 mA rechargeable batteries. We used our agent based directed diffusion protocol (Shakshuki et al. 2008) to diffuse the query into the network. For our experiments, we aggressively sampled for 3 h, every second all the 17 sensory attributes from each sensor mote to produce a sensory trace at BS. These traces acted as historical data for our methodology. We combined the historical trace logs for all motes and then applied our methodology to find the important attributes along non correlated PC. The goal of our case study is to thoroughly examine the following three research questions:

Q1. How effective is our methodology to reduce the energy consumption of WSN?

Our first experiment was conducted to compare the energy consumptions of WSN using a query containing all the attributes with that of an optimal query that only consists of top attribute from each PC. The batteries were fully charged before the start of experiment and motes were laid out carefully in a grid layout, i.e., exactly one meter apart from each other. Both (standard and optimized) queries were exposed to WSN for 150 min. The experiment was repeated fivefolds and environment was kept constant. Figure 3 shows the average result of fivefolds. Optimal query constructed using our methodology clearly banks 50% more energy. Figure 6 shows the consumption of optimal query when sensor is deployed in different layout. As expected, grid consumes the least energy, whereas random deployment is energy expensive.

Q2. What affect our methodology has on the health of WSN?

We conducted our second experiment to find out how quickly sensor motes expire (health) in WSN using both standard and optimized query. We powered-off the radio

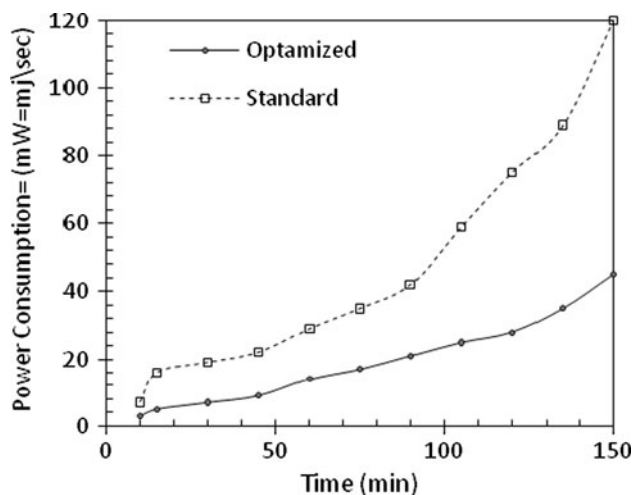


Fig. 3 Energy consumption

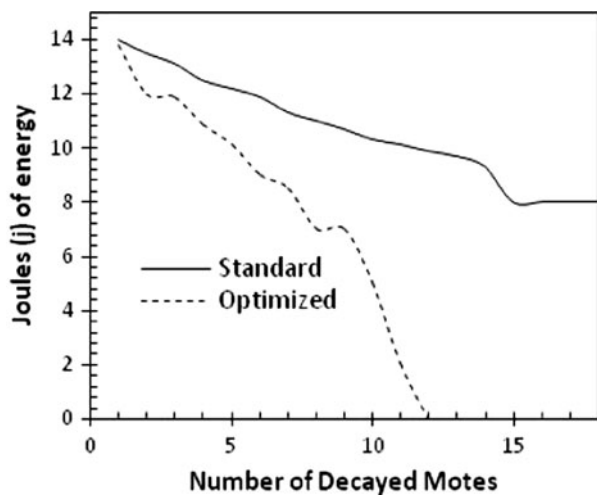


Fig. 4 Comparison of node decay

for each mote if it consumed more than 0.7 J of energy and consider it expired. We repeated the experiment five times. Figure 4 reports the average results of all fivefolds. By using optimal query, the health of WSN is found 53% better as compared to standard query. We find sudden drop in node decay in standard query once 50% of nodes expire. On close examination of network, it is revealed that due to expired motes, communication holes are created in network due to the aggressive use of optimal routing path to BS. Hence, live motes need to route the data through non optimal path, thereby depleting the energy rapidly.

Q3. Does cycling the correlated attribute in optimal query further improve the energy consumption of motes?

The motivation behind this research question is that each sensor on a mote drains specific energy from batteries. Therefore, cycling the top correlated attributes will cause an even energy depletion of mote's energy.

Attribute cycling in an optimal query also decreasing the biasing of reproduced sensor data based on only one fixed query attribute. This also reduces the skew in reproduced sensor data by latent corner cases which slipped through the association rule mining. We conducted an experiment where we cycled the Top_k attributes from each PC after 30, 60, 90, 120, 150 s. Figure 5 shows that too frequent switching of attributes in an optimal query is not required. Each change in query has a certain underlying energy cost associate for its diffusion into a sensor network. This is due to the fact that for each diffused query, the sensor motes have to create the models of its neighboring motes from whom the query was received. The models of neighbor motes help a mote to route the data back to the sink when an event is sensed. The detailed underlying energy cost associated with model construction can be found in our previous research work (Shakshuki et al. 2008).

6 Performance evaluation

To evaluate the effectiveness of our methodology, we used two metrics:

1. *Network lifetime* The network lifetime is defined as the time duration before the first sensor node runs out of power.
2. *Accuracy* We define accuracy as the ability of our methodology to correctly synthesize all the sensory values of a mote from the reduced set of query attributes reported at BS by optimal query. It is the ratio between the correctly synthesized attributes values and that of the actual attributes value reported by standard query.

The result of our methodology is accurate for a given round, if it correctly synthesizes all the motes sensory data for the respective round, i.e., for the “loop” attribute of the query as shown in Table 1. For example, [A, B, C, D, E] are the sensors mounted on the mote. Our methodology DR technique suggested using only sensor A and D to query the network. At the BS, our methodology synthesized the value of sensor [B, C and E] with 90% confidence level and achieved more than 90% accuracy. We measure the accuracy of WSN using the following definition:

Let the total number of sensor mounted on the mote is (T). The number of sensory attributed used in a query (Q) is (A). The number of sensory attributes (S) required to synthesized at BS are $S = (T - A)$.

$$\text{Acc} = \frac{C}{S} \quad (1)$$

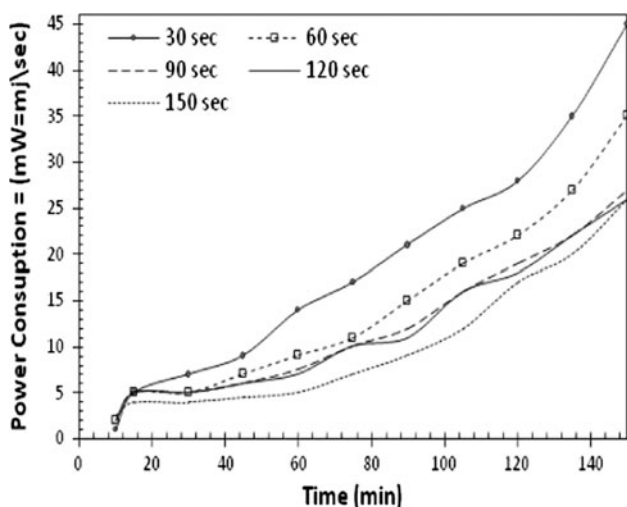


Fig. 5 Cycle interval

$$Acc_round = \frac{1}{M} \cdot \sum_{i=1}^M (Acc_i) \tag{2}$$

$$Acc_network = \frac{1}{R} \cdot \sum_{i=1}^R (Acc_round_i) \tag{3}$$

where, Acc is accuracy for a mote for a round and C is the correct number of sensory attributes synthesized by our methodology. Acc_round is the accuracy of all the motes M in a WSN for a data round. Acc_network is the accuracy achieved by our methodology for all the data rounds R satisfying the WSN query (Q).

Approach: To evaluate the performance of our methodology based on the two metrics defined, we conducted a set of experiments by deploying two WSN, A and B spanning over nine zones of a botanical garden. Each WSN consisted of nine motes and a BS. Every zone was equipped with two sensor motes. Each mote in a zone serviced different query and sent sensory data back to its corresponding WSN’s BS. A standard query requesting motes to report all of their data of their sensory attributes to the base station over the period of 60 h was diffused into WSN A. Optimized query requesting reduced set of attributes (only 2) as suggested by the DR technique employed by our methodology was injected into WSN B. All the motes were equipped with alkaline long life AA batteries of Energizer company having average 1.225 V during discharge, 2,100 mA h, i.e., 2.69 W h (9,360 J) of energy. During the experiment we avoided the use of three LED (blue green and red) on mica motes as both queries have different numbers of sensory attributes, which results into different blinking operation for each query, based on the type of sensor actuated form sensing on the mote. Since the blinking operation of the LED drains lot of current, i.e., from 3.3 to 5.3 mA and only serves as operational

visualization for a mote. Avoiding the using of LED helps us to produce unbiased comparison for energy consumption of WSN A and B.

The TR1000 radio transceiver module for wireless communication notoriously exploit the energy consumption by withdrawing around 20.5 mA current while transmitting data, i.e., RX transmission with no LED on. We marked the mote expired after reaching 2 V. At this stage, even the 1.8 V output internal regulator of the mote cannot guarantee the stable power consumption for the radio thereby affecting the ability of radio to transmit properly (Fig. 6).

Findings: Figure 7 shows that in WSN, which is exposed to standard query (requesting the data of all sensory attributes) has half the life time as compared to optimized query constructed on the recommendation of our methodology. The WSN is able to handle 3,600 data rounds before a mote expires. Moreover, the used of optimized query ensure and even burnout of WSN energy, avoiding sudden communication holes in network due to expired nodes as shown in Fig. 8. Communication holes in WSN are created when intermediate node(s) relaying the sensed data from other nodes to base station expires. The surrounding nodes (neighbors of the expired node) then have to use intermediate nodes among non optimal routing path to send sensed data to base station. In worst case, they need to make direct long range transmissions to the BS resulting into hasty consumption of battery power (Heinzelman et al. 2000).

Figure 8 is a bar plot for the average energy consumed per 60 rounds by the standard and optimized query. Since in standard query all the sensors on the mote are utilized to sample the environment (analog signal), the analog to digital conversion (ADC) channel draws more current (around 0.5 mA) as compared to motes exposed to optimized query. Hence, the batteries drain out rapidly. We can see that between 900 and 960 round bucket, the energy consumption of the standard query is at the peak. This is due to the reason that three of the motes near the base station are expired and remaining 6 motes are making direct transitions to the base station. The energy consumed in transmitting over large distance is proportional to distance (Heinzelman et al. 2000). After the 960th round, we see the average energy consumed per 60 bucket starts to decrease and by 1,200 round it becomes steady. This is due to the underlying reason that after the 960th round the nodes making direct transmission to the base station slowly start to expire and average energy consumption per bucket decreases. By 1200th round there is only two nodes left making direct transmissions.

We calculated the accuracy of our methodology using the definitions in Eqs. 1–3. Our methodology is able to achieve 93% accuracy as shown in Figs. 9 and 10. Accuracy of our methodology under worst case scenario is

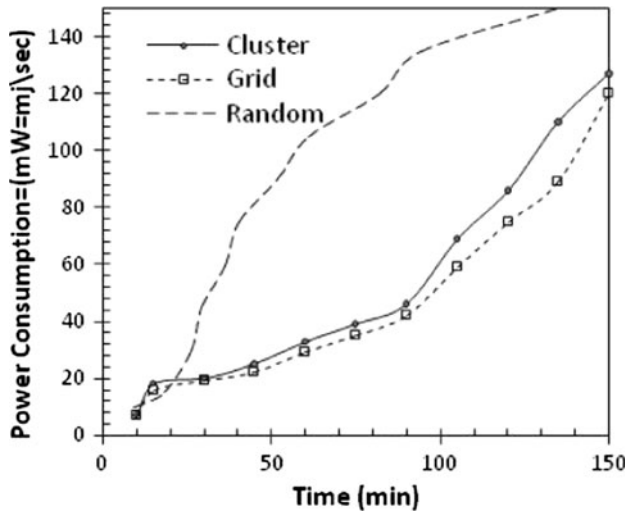


Fig. 6 Energy consumption for different layouts

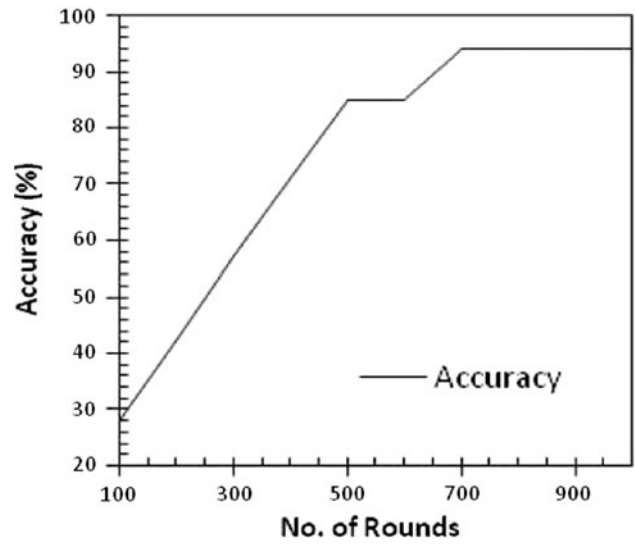


Fig. 9 Relationship between accuracy and rounds

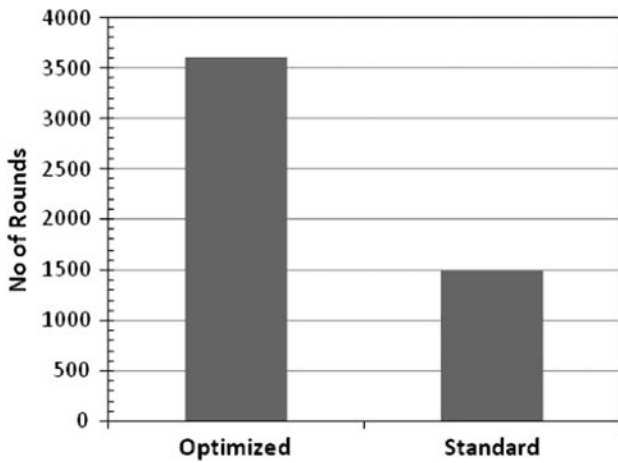


Fig. 7 network life time

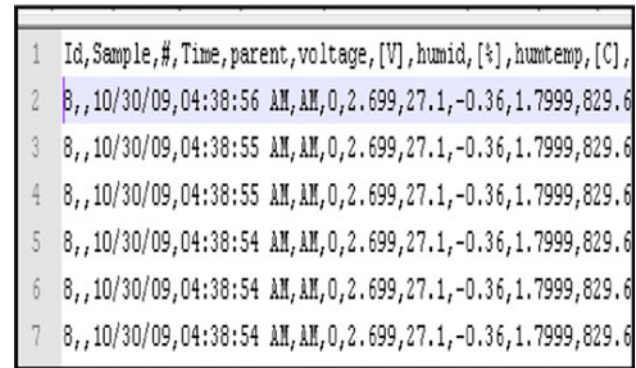


Fig. 10 Sample log trace of a sensor mote

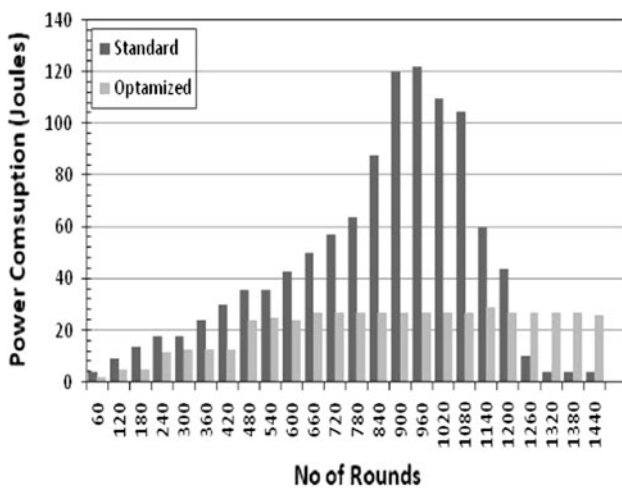


Fig. 8 Average energy consumption per 60 rounds

shown in Fig. 11a. A query was diffused into the WSN requesting notes to report the value of the required attributes every minute for next 1,440 rounds, i.e., for 24 h. We randomly picked 200 samples, i.e., data rounds using out-of-bag (OBG) approach to ensure no two samples are selected twice. Apriori algorithm was utilized to construct association rules on the 200 historical samples (rounds). We used WSN A and B to diffuse optimized and standard query. At the base station the data reported by optimized query (containing reduced set of sensory attributes, i.e., 2 attributes) was synthesized to obtain the values of all the sensory attributes using the association rules. The result of each mote for the 500 round was compared with the actual attributes value reported by the notes exposed to standard query. Figure 11a shows that there is 100% accuracy achieved by our methodology for the first 450 rounds of data, i.e., for the first 7.5 h of sensory data collected for the optimal query. After passing 450th round, the accuracy of our methodology destabilizes. Figure 11b is the line plot of

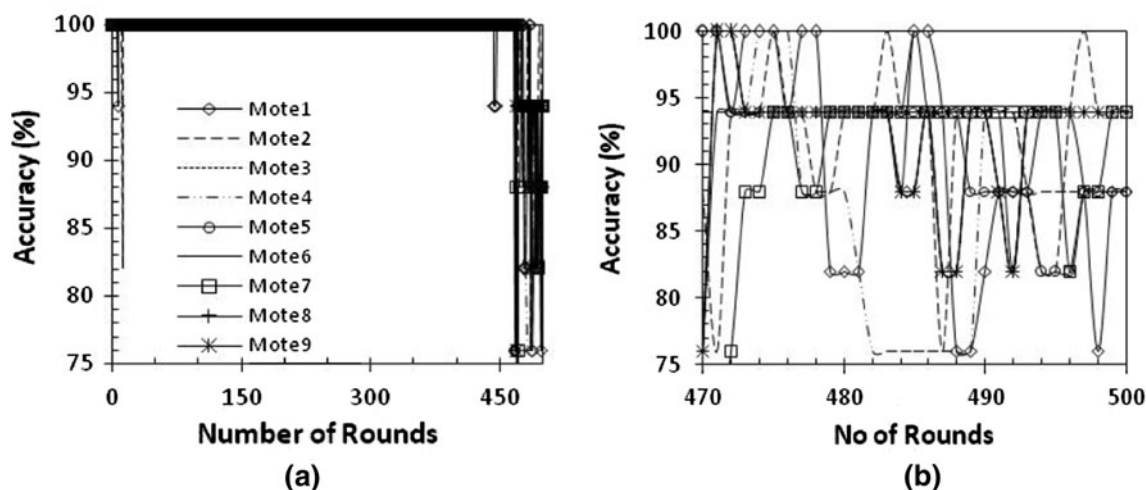


Fig. 11 Accuracy per mote

accuracy from all the motes from 470 rounds to 500 round. Surprisingly, there is no clear trend and accuracy of our methodology for each mote fluctuates for every round. On close examination of the rules and sensory attribute values obtained at sink after 450th round, it was found that the rules accommodating photonic and humidity had less than 50% of confidence level for the synthesized data at the BS. All the 200 historical samples to construct rules were the rounds reported to the BS at night time. For our optimal query, these rules had failed to synthesize all the sensory data with 90% confidence level as the 450th round occurred at 6:00 am (morning) and the correlation between the photonic and humidity sensor changed by them, in contrast to the night time.

We did an experiment to find out what is minimal sample size based on data round per minute to achieve 90% of accuracy. We varied the historical sample size (rounds) for construction of association rules. Figure 8 shows that around 800 samples are enough to achieve 90% of accuracy.

7 Conclusions and future work

A novel data reduction method is proposed to optimize the execution of periodical query based on historical sensor data. The reduced set of data reported at BS is synthesized using Apriori algorithm to construct complete data. The methodology is explained in action using the PCA. The proposed methodology helps to construct optimal query that prolongs the life time of WSN by 50% as compared to standard query. Experiments are conducted to validate the effectiveness of methodology. Although the methodology is currently applicable to applications where data are not aggregated and compressed, in future we want to employ

techniques such as liner regression and maximum likelihood estimation to construct association rules for aggregated data.

References

- Bajaber F, Awan I (2010) Energy efficient clustering protocol to enhance lifetime of wireless sensor network. *J Ambient Intell Human Comput* 1(4):239–248
- Beyer K, Goldstein J, Ramakrishnan R, and Shaft U (1999) When is “nearest neighbor” meaningful? *Lecture notes in computer science*, vol 1540, pp 217–235
- Bonnet P, Gehrke J, and Seshadri P (2001) Towards sensor database systems. In: *Proceedings of the 2nd international conference on mobile data management*, vol 43, pp 551–558
- Box MJ, Box RM (1996) Computation of the variance ratio distribution. *Comput J* 12(3):277–278
- Bronnimann H, Chen B, Dash M, Haas P, Scheuermann P (2003) Efficient data reduction with EASE. In: *Proceedings of the 9th ACM international conference on knowledge discovery and data mining*, pp 59–68
- Chen M, Fowler ML (2004) Data compression trade-offs in sensor networks. In: *Proceedings of mathematics of data/image coding, compression, and encryption VII, with applications*, vol 5561, pp 96–107
- Deshpande A, Guestrin C, Madden S, Hellerstein J, Hong W (2004) Model-driven data acquisition in sensor networks. In: *Proceedings of the 13th international conference on very large databases*, vol 3, pp 588–599
- DIM Project (2007) DIN for sensor networks. <http://enl.usc.edu/projects/dim/index.html>. Accessed 7 May 2007
- Heinzelman WR, Chandrakasan A, Balakrishnan H (2000) Energy-efficient communication protocol for wireless micro sensor networks. In: *Proceedings of the 33rd Hawaii international conference on system sciences*, 4–7 Jan 2000, vol 2
- Hiyama M, Ikeda M, Barolli L, Takizawa M (2010) Performance analysis of multi-hop ad-hoc network using multi-flow traffic for indoor scenarios. *J Ambient Intell Human Comput* 1(4):283–293
- Jolliffe IT (2002) *Principal component analysis*, 2nd edn. Springer, New York

- Ju H, Cui L (2005) EasiPC: a packet compression mechanism for embedded WSN. In: Proceedings of the 11th IEEE international conference on embedded and real-time computing systems and applications, pp 394–399
- Kaiser F (1974) Factorial simplicity index and transformation. *Psychometrika* 49(2):277–295
- Le S, Julie J, Francois H (2008) FactoMine: an R package for multivariate analysis. *J Stat Softw* 25(1):1–18
- Liu X, Ma Z, Cao X, Liu S (2010) A query processing method in multi-user scenarios. In: proceedings of 5th international ICST conference on communication and networking, pp 25–33
- Madden S, Franklin MJ (2002) Fjording the stream: architecture for queries over streaming sensor data. In: Proceedings of the 18th international conference on data engineering, p 555
- Madden S, Hellerstein JM (2002) Distributing queries over low-power wireless sensor networks. In: Proceedings of the 2002 ACM SIGMOD international conference on management of data, 2–6 Jan 2002
- Naoto K, Shahram L (2005) A survey on data compression in wireless sensor networks. In: Proceedings of the international conference on information technology: coding and computing, vol 02, pp 8–13
- Nose Y, Kanzaki A, Hara T, Nishio S (2010) A route construction based on measured characteristics of radio propagation in wireless sensor networks. *J Ambient Intell Human Comput* 1(4):259–270
- Ozsy MT, Valduriez P (1991) Principles of distributed database systems. Prentice Hall, Englewood Cliffs
- Raghunathan V, Schurgers C, Park S, Srivastava MB (2002) Energy-aware wireless microsensor networks. *IEEE Signal Process Mag* 19(2):40–50
- Rakesh A, Ramakrishnan S (1994) Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th international conference on very large databases, pp 487–499
- Shakshuki E, Malik H, Denko M (2008) Software agent-based directed diffusion in wireless sensor network. *J Telecommun Syst* 38(4):161–174
- Shatdal A, Naughton JF (1995) Adaptive parallel aggregation algorithms. In: Proceedings of the 1995 ACM SIGMOD international conference on Management of data, pp 104–114
- Silberstein A, Braynard R, Ellis C, Munagala K (2006) A sampling-based approach to optimizing top- k queries in sensor networks. In: Proceedings of the 22nd international conference on data engineering, p 68
- Tik LB, Khuan CT, Palaniappan S (2009) Monitoring of an aeroponic greenhouse with a sensor network. *Int J Comput Sci Netw Secur* 9(3):240–246
- Uchida N, Takahata K, Shibata Y (2010) Transmission control methods with multihopped environments in cognitive wireless networks. *J Ambient Intell Human Comput* 1(4):249–257
- Wu M, Xu J, Tang X, Lee C (2007) Top- k monitoring in wireless sensor networks. *IEEE Trans Knowl Data Eng* 19(7):962–976
- Xu R, Li Z, Wang C, Ni P (2003) Impact of data compression on energy consumption of wireless-networked handheld devices. In: Proceedings of 23rd international conference on distributed computing systems, pp 302–311
- Yazti Z, Vagena Z, Gunopulos D, Kalogeraki V, Tsotras V (2005) The threshold join algorithm for top- k queries in distributed sensor networks. In: Proceedings of the 2nd international workshop on data management for sensor networks, pp 61–66
- Yeo M, Seong D, Yoo J (2009) Data-aware top- k monitoring in wireless sensor networks. In: Radio and wireless symposium, 18–22 Jan 2009, pp 103–106
- Yu CT, Chang CC (2000) The state of the art in distributed query processing. *ACM Comput Surv* 16(4):399–433
- Zhang Y, Hull B, Balakrishnan H, Madden S (2007) ICEDB: intermittently connected continuous query processing. In: Proceedings of international conference on data engineering