Approximate Querying in Wireless Sensor Networks

Yuzhen Liu Weifa Liang
Department of Computer Science
Australian National University
Canberra, ACT 0200, Australia

***Abstract*— In this paper, we study the maximization problem of network lifetime for answering a sequence of aggregate queries based on snapshot data. We build a series of nearly optimal representative routing trees for query evaluation, where a representative routing tree is such a tree rooted at the base station that each node in it represents a set of non-tree nodes by holding their historical data (snapshot data). A representative routing tree is optimal if the minimum residual energy among its nodes is maximized, and the number of nodes in the tree is minimized. Due to the unpredictability of future queries, we will focus on the construction of individual optimal representative routing trees in order to solve the maximization problem of network lifetime. We first show the optimal representative routing tree problem is NP-complete. Instead, we then devise two heuristic algorithms for it. We finally conduct extensive experiments by simulations to evaluate the performance of the proposed algorithms in terms of the network lifetime and the average size of representative routing trees. The experimental results showed the proposed algorithms outperform an existing algorithm significantly.**

Keywords: *Wireless sensor network, aggregate query evaluation, network lifetime, snapshot-based query, correlated data gathering, representative routing tree.*

I. Introduction

Sensor networks have been used for environmental monitoring purposes [10]. During this course, the large volume of sensed data generated by sensors is needed to be either collected at the base station for further processing or processed within the network to answer user queries. The sensor network thus has also been treated as a virtual database by the database community [9], which is essentially different from the traditional databases. It is impossible sometimes to store all sensed data in a central site (the base station) simply because the data generated by sensors are continuing data streams, and the volume of the data is so huge. On the other hand, the battery-powered sensors will quickly become inoperative due to large quantity energy consumption if all sensed data must be sent to the base station. Particularly, energy conservation in the operations of sensor networks is of paramount importance, which poses great challenges on the evaluation of queries in such a database efficiently and effectively.

To reduce the energy consumption in sensor networks, instead of providing the exact answer for each user query by all sensors participation [7], sometimes an approximate solution is acceptable, which is obtained by using a subset of sensors in the network [13]. Unlike the previous work that focused on approximation solutions for aggregate queries in sensor networks, two model-driven approaches are recently proposed [2], [5]. Deshpande *et al* [2] introduced a model-driven data acquisition framework, which uses a trained statistical model at the base station instead of in-network query processing to answer user queries with high confidence. Thus, the energy at sensors is saved through limiting the number of sensor readings. It is assumed in their model that the base station is a powerful computer that has unlimited power supply. In contrast to this model, Kotidis [5] proposed another model-driven query processing approach, referred to as the *snapshot-based* query processing. Essentially, this two-stage approach partitions the sensor nodes in a sensor network into two categories: the *representative nodes* and the *represented nodes*. In the first stage, a certain number of sensors are chosen as the representatives of those unchosen sensors. Each chosen sensor will represent its members solely. None of the members can be represented by two representative nodes. Each represented member usually is at sleep mode but periodically sends its heart-beat signal and sensed data to its representative node. Given two neighboring nodes, which one represents the other is determined by a given correlation metric. For a represented node, its representative node will provide an estimate of its actual value using a formula, based on the historical data (snapshot data) it sent to the representative node [5]. In the second stage, a spanning tree rooted at the base station spanning all the nodes is first built using a Breadth-First-Search approach in [8], [9]. A routing tree is then obtained by pruning branches of the spanning tree until each leaf node in the resulted tree is a representative node. Note that in this routing tree, some non-representative nodes that serve as relay nodes for others are also included. Each representative node in the routing tree reports its own and its represented nodes' results to the base station through relay nodes. In the rest of this paper, this approach is referred as to algorithm `snapshot`. However, algorithm `snapshot` has its limitations. To respond to each incoming query, a dedicated routing tree for the query as above must be built. Consequently, the energy overhead on frequently routing tree building cannot be ignored. On the other hand, despite the number of representative nodes chosen in the first stage may be small, it cannot be guaranteed that the number of nodes in the routing tree is small as well, the tree may contain much more relay nodes which are not the representative nodes. This implies that only very few nodes in the tree will become the represented nodes.

Motivated by the work due to Kotidis [5], in this paper we study the maximization problem of network lifetime for answering a sequence of unknown aggregate queries based on the snapshot data. We build a series of snapshot-based representative routing trees for query evaluation. Although Kotidis [5] also uses the snapshot-based representative tree concept, there are some essential differences between his work

978-1-4244-2020-9/08/$25.00 ©2008 IEEE

and ours. (i) We introduce not only the correlation metric d as he did in [5], but also the correlation threshold θ that models the query quality requirement by different users. The experimental results show that there is a non-trivial trade-off between the quality of the query result and the amount of energy consumption per query. (ii) During the construction of a representative routing tree, on one hand, we aim to minimize the number of nodes in the tree. On the other hand, we also require that the tree serve for query evaluation as long as possible. A node is chosen to be included in the tree after having taken into account not only how many other nodes it can represent but also how much its current residual energy is. Two heuristics are proposed to trade-off these two factors. In his algorithm, the procedures of choosing the representative nodes and constructing the routing tree are separate, which may result in lots of non-representative nodes included in the routing tree, whereas we deal with these two procedures jointly. As a result, the number of nodes in our routing tree is controlled. (iii) Given a sequence of unknown aggregate queries, our optimization objective is to maximize the network lifetime by answering as many queries in the sequence as possible, while each query in the sequence will be evaluated by only one of the nearly optimal representative routing trees. The nodes in different trees may not be different, thereby, the network lifetime is prolonged. In his algorithm, the major motivation is to build the snapshot based data acquisition model and the author focused on a single query by building a routing tree such that the number of nodes in the tree is as small as possible.

II. Preliminaries

A wireless sensor network can be modelled by an *undirected graph* $G = (V, E)$, where V is the set of sensor nodes, and there is an undirected edge (u, v) in E if and only if nodes u and v are within the *transmission range* r_t of each other, $|V| = n$ and $|E| = m$. Among the n sensors, one is the base station, which has unlimited energy supply. Each sensor equipped with an omni-directional antenna can monitor its vicinity. The *sensing area* by a sensor is a circle with radius r_s, which is referred to as *sensing range*. Each node consumes its energy when performing message transmissions, receiving messages, sensing, and computing. The dominant energy consumption at a node is its transmission energy consumption, followed by the reception energy consumption.

Two neighboring nodes u and v in V are *highly correlated* if $(u, v) \in E$ and $d(u, v) \geq \theta$, given a wireless sensor network $G(V, E)$, data correlation metric d, and data correlation threshold θ, $0 < \theta < 1$. For example, if we take the metric d as the ratio of their overlapping area to the sensing area of a sensor, then, the metric says that two neighboring nodes are highly correlated if the ratio between them is at least θ. Node u can be *represented* by node v if they are highly correlated. *A representative routing tree* in a wireless sensor network $G(V, E)$ is a tree rooted at the base station such that every non-tree node is represented by a tree node, i.e., for each node $v \in V$, either v is in the tree or v is highly correlated with at least one tree node. *An optimal representative routing tree* in G is a representative routing tree such that (i) the number of nodes in the tree is minimized; and (ii) the minimum residual energy among the tree nodes is maximized. Here (i) implies that more representative routing trees can be built if each of the trees contains fewer nodes, and (ii) ensures that each tree can survive as long as possible. The optimal representative routing tree will be used for the evaluation of aggregate queries (issued at the base station) by providing approximate answers, based on the snapshot data of represented nodes stored at their representative nodes. The lifetime of an optimal representative routing tree is determined by the number of queries it evaluated and defined by the time when the first node in the tree dies. The *maximization problem of network lifetime* is defined as follows. Given a sensor network $G(V, E)$ with a base station that has unlimited energy supply and all the other nodes have the same initial energy capacity IE, sensing range r_s and transmission range r_t, assume that there is a sequence of unknown aggregate queries and queries arrive and are evaluated one by one by the system. The problem is to construct a series of optimal representative routing trees such that the sum of lifetime of the optimal representative routing trees is maximized.

In this paper, we aim to build a series of optimal representative routing trees to maximize the sum of lifetime of the representative routing trees, thereby maximizing the network lifetime. In the rest of this paper we shall focus only on the optimal representative routing tree problem.

III. Algorithms for Optimal Representative Routing Trees

A. NP Hardness

Lemma 1: The optimal representative routing tree problem in a wireless sensor network $G(V, E)$ is NP-Complete.

Proof: We first reduce the *minimum connected dominant set* (MCDS for short) problem to the optimal representative routing tree problem within polynomial time as follows.

Given an instance $G(V, E)$ of MCDS and an integer K, the decision version of MCDS is to find a node subset S of size $|S| \leq K$ such that the subgraph induced by the nodes in S is connected, and for every other node $v \in V - S$ there must have an edge $(v, u) \in E$ where $u \in S$. The problem is well known to be NP-Complete [3].

Given the above instance $G(V, E)$ of MCDS and K, a wireless sensor network instance $M = (N, L)$ of the optimal representative routing tree problem is constructed as follows. $N = V$. If there is an edge $(u, v) \in E$, then the corresponding two neighboring sensor nodes u and v are highly correlated and there is an edge $(u, v) \in L$. For simplicity, we assume that all sensor nodes have identical residual energy. Thus, the decision version of the optimal representative routing tree problem in $M(N, L)$ is to find a representative routing tree T such that the number of nodes in the tree is no greater than K. Let $T(V)$ be the set of nodes in T. Then, a solution $T(V)$ for the latter problem is a solution for MCDS, because $|T(V)| \leq K$ and all the nodes in $T(V)$ is connected, and for every $u \notin T(V)$, there must be an edge $(u, v) \in L$ between u and v, where $v \in T(V)$.

Given a tree, whether it is an optimal representative tree can be verified within polynomial time. Thus, the optimal representative routing tree problem is NP-Complete. ■

B. Heuristic algorithms

Due to the NP hardness of the optimal representative routing tree problem, in the following we focus on devising two heuristic algorithms for it.

1) *Overview of heuristics* We assume that the entire network lifetime consists of several stages. At each stage a representative routing tree will be used. Thus, the number of representative routing trees corresponds to the number of different stages of the network lifetime. Within a stage, all the queries posed during that period will be evaluated by the corresponding representative routing tree. A shift from the current stage to the next stage means that a new routing tree based on the current residual energy of each node will be built, and it will be employed at the next stage. In the rest, we focus on building a representative routing tree after taking into account the spatial data correlation, based on the snapshot data at those represented nodes.

2) *Heuristic algorithm one* The heuristic algorithm is as follows. The node set V is partitioned into three subsets, $T(V)$, $C(V)$, and U, where $T(V)$ is the set of tree nodes, $C(V)$ is the set of correlated nodes that are represented by the tree nodes already, and U is the set of nodes that are neither in the tree nor being correlated with any tree nodes. $V = T(V) \cup C(V) \cup U$.

The representative routing tree is built greedily. Initially, the representative routing tree contains the base station only. Each time a node $u \in U$ is chosen to be added to the tree if there is an edge between a tree node and u and $h_1(u)$ is minimized, where $h_1(u)$ for $u \in U$ is defined as follows.

$$h_1(u) = \frac{IE - RE(u)}{N'(u)}, \tag{1}$$

where $N'(u)$ is the number of those unrepresented neighboring nodes of u that are highly correlated with u under the metrics d and θ, IE is the initial energy capacity, and $RE(v)$ is the residual energy at node v. Once u is chosen, all the nodes in $N'(u)$ will be added to $C(V)$, and $U = U - N'(u) - \{u\}$. Otherwise (no such $u \in U$ exists), a node from $C(V)$ is chosen and added to the tree using the same metric as (1).

It can be observed from the definition of h_1, it favors the node with high residual energy and large number of highly correlated, unrepresented neighboring nodes. If two non-tree nodes have the same residual energy, the node that is highly correlated with more unrepresented nodes will be selected and added to the tree. On the other hand, if two non-tree nodes represent the same number of unrepresented nodes, the heuristic will select the node with more residual energy and add it to the tree. The detailed algorithm is given below.

Algorithm `Representative_Routing_Tree`
Input: $G(V, E)$, r_s, RE(), d, θ
Output: A representative routing tree $T(V', E')$, $V' \subseteq V$ if it exists.
begin
1. $T(V) \leftarrow \{s\}$; /* s is the base station */
2. $C(V) \leftarrow \emptyset$;
3. $U \leftarrow V - \{s\}$;
4. **while** $(U \neq \emptyset)$ **do**
5. **if** there is a node $u_0 \in U$ such that $\frac{IE-RE(u_0)}{N'(u_0)} = \min_{(v \in T(V), u \in U, (u,v) \in E)}\{\frac{IE-RE(u)}{N'(u)}\}$ /* let $(u_0, v_0) \in E$ be the edge and $v_0 \in T(V)$ */
6. **then** $T(V) \leftarrow T(V) \cup \{u_0\}$;
7. $C(V) \leftarrow C(V) \cup N'(u_0)$;
8. $parent(u_0) \leftarrow v_0$; /* set the parent of u_0 in the tree */
9. $U \leftarrow U - N'(u_0) - \{u_0\}$;
10. **else if** there is a node $u_1 \in C(V)$ such that $\frac{IE-RE(u_1)}{N'(u_1)} = \min_{\{v \in T(V), u \in C(V), (u,v) \in E\}}\{\frac{IE-RE(u)}{N'(u)}\}$ /* let $(u_1, v_1) \in E$ be the edge and $v_1 \in T(V)$ */
11. **then** $T(V) \leftarrow T(V) \cup \{u_1\}$;
12. $parent(u_1) \leftarrow v_1$;
13. $C(V) \leftarrow C(V) \cup N'(u_1) - \{u_1\}$;
14. $U \leftarrow U - N'(u_1)$;
else exit; /* the tree does not exist; */
endif;
endif;
endwhile;
15. $V' = T(V)$;
16. $E' = \{\langle v, u \rangle | u \in T(V), v \in T(V)$ *and* $parent(u)$ *is* $v\}$;
17. return $T(V', E')$;
end.

We have the following lemma.

Lemma 2: Given a wireless sensor network $G(V, E)$ with correlation metric d and correlation threshold θ, $0 < \theta < 1$, the solution $T(V', E')$ delivered by algorithm `Representative_Routing_Tree` is a representative routing tree of G.

Proof: It can be seen from the construction of T that T is tree. We now show that for any node $v \in V$, either v is in $T(V)$ or v is represented by a node in $T(V)$ by contradiction. Assume that there is such a node $v \in V$ that neither it is in $T(V)$ nor none of the nodes in $T(V)$ can represent it. That is, v is in neither $T(V)$ nor $C(V)$. Since $V = T(V) \cup C(V) \cup U$, we have $U \neq \emptyset$, and the algorithm will not terminate, which means that the construction of T has not finished yet. This contradicts that T is the final solution delivered by the algorithm. The lemma thus follows. ■

Theorem 1: Given a wireless sensor network $G(V, E)$ with a correlated metric d and correlation threshold θ, $0 < \theta < 1$, there is a heuristic algorithm for the optimal representative routing tree problem, which takes $O(mn)$ time, where $n = |V|$ and $m = |E|$.

Proof: Following Lemma 2, the tree delivered by the proposed algorithm is a representative routing tree.The time complexity of the proposed algorithm is analyzed as follows.

Step 1 to Step 3 take $O(1)$ time. Step 5 takes $O(m)$ time. Step 6 to Step 9 take $O(n)$ times at most. Step 10 takes $O(m)$ time. Step 11 to Step 14 take $O(n)$ time. The **while** loop is executed at most n times. Thus, the total running time for Step 5 to Step 14 is $O(mn)$ time. Step 15 and Step 16 take $O(n)$ time. Algorithm `Representative_Routing_Tree` thus takes $O(mn)$ time. ■

3) *Heuristic algorithm two* The previous heuristic implies that there is a non-trivial tradeoff between the number of representative routing trees and the lifetime of each of the trees. Motivated by this observation, we aim to design a new heuristic that maximizes not only the residual energy at each representative node but also the number of nodes being represented by each representative node.

It has been shown that a heuristic based on the exponential function of energy utilization at nodes is very useful in the design of on-line algorithms for unicasting and multicasting in ad hoc networks [4], [6]. Here, a heuristic function based on the exponential function of energy utilization for the optimal representative routing tree problem is given as follows.

$$h_2(u) = \frac{N'(u)}{IE(\lambda^{\beta(u)} - 1)}, \quad (2)$$

where $\lambda > 1$ is a constant, and $\beta(u) = \frac{IE - RE(u)}{IE}$ is the energy utilization ratio at u so far.

The heuristic is to maximize the value of $h_2(u)$. It can be seen from (2) that this heuristic penalties those nodes that have small percentages of residual energy with increase of the value of λ heavily. On the other hand, it favors choosing a node as the representative node if it can represent as many other nodes as possible. The algorithm for building the representative routing tree by employing heuristic function h_2 is similar to algorithm `Representative_Routing_Tree`. The only difference is that the node with the maximum value of h_2 is chosen into the tree in Step 5 and Step 10 in algorithm `Representative_Routing_Tree`.

4) *Network lifetime maximization algorithm* The maximization problem of network lifetime is then solved by calling procedure `Representative_Routing_Tree` iteratively, which is detailed as follows. A representative routing tree based on the initial energy at nodes is built to accommodate the queries during the living period of the tree until the first node in the tree dies due to the expiration of its energy. The next representative routing tree is then constructed based on the current residual energy among the nodes to respond to subsequent queries. The procedure is repeated until no further representative routing tree can be built based on the current residual energy. The algorithm for the maximization problem of network lifetime can be described as follows.

Algorithm `Network_Lifetime_Maximization`
Input: $G(V, E)$, r_s, RE(), d, θ
Output: The network lifetime.
begin
1. $lifetime \leftarrow 0$; /* the network lifetime*/
2. **while** $true$ do
3. **call** `Representative_Routing_Tree`;
4. **if** a representative routing tree exists
5. **then** $lifetime \leftarrow lifetime + lifetime1$;
   /* *lifetime1* is the lifetime delivered by algorithm */
   /* `Representative_Routing_Tree` */
   else exit;
   endif;
   endwhile;
6. return $lifetime$;
end.

For the sake of convenience, algorithm `Network_Lifetime_Maximization` employing heuristic function h_1 and h_2 is referred to as `NLM1` and `NLM2`, respectively.

IV. Potential Applications

The representative routing tree can be used to collect training data from the network to build global query models at the base station. These models will later be used to answer user queries directly, rather than to get the query result through in-network processing. For example, in Deshpande *et al* model [2], the initial training data set must be sent to the base station, the representative routing tree can be used for such a purpose. The representative routing tree can also be used for collecting the updated data from some sources if there is any significantly updating among the sensor nodes. Recently Meliou *et al* [11] proposed a strategy for forwarding the updating sources data to the base station by finding a Eulerian-tour including all sources, which is limited to the case where the number of nodes involved is small and the message forwarded by each node in the tour is fixed. Otherwise, the cost is expensive and a representative routing tree for such collection purpose would be better.

Although the representative routing tree is used for approximate querying, it can be applicable for other query applications. For example, it can be used to answer a class of spatial-correlated aggregate queries approximately. Due to the fact that in most cases the sensors are deployed randomly by the aircraft, some area within a monitored region may have lots of redundant sensors. This will result in the highly correlated data among nearby sensors. To prolong the network lifetime by exploiting sensor correlation, von Rickenbach and Wattenhofer [12], [1] devised an algorithm for data gathering by building a spanning routing tree rooted at the base station. The sensed data by each sensor is compressed through internal nodes in the tree before it is transferred to the base station. From this scenario, it can be seen that the sensed data in nearby sensors are highly correlated. To alleviate the volume of transmitted data at each node, it is necessary to remove the correlated data before the compression. However, if an approximate result is acceptable, instead of building a spanning tree for evaluating queries, a representative routing tree suffices, which consists only of the representative nodes, where a representative node is chosen from these highly correlated nodes and represents them. The same compression technique can then be applied to the internal nodes in the representative routing tree. Thus, the energy consumption at each node is significantly reduced, since a fewer number of sensor nodes are included in the representative routing tree and less computing time at each node for correlation data is required. In addition, those sensor nodes that have not been included in the current representative routing tree can be used in later routing trees. Thus, the network lifetime will be substantially prolonged.

In the following we use an example to demonstrate that the representative routing tree can be stand alone and used to evaluate typical aggregate queries in databases by returning approximate results.

Theorem 2: Given a wireless sensor network and approximation error ϵ, there is an approximation solution for queries like `MAX, MIN, AVG`, using the representing routing tree. In other words, let y be the exact value of one of the functions of `MAX, MIN, AVG` and $\bar{y}$ the approximate value returned by the proposed algorithm, then $|y - \bar{y}| \leq \epsilon$.

Proof: We here only consider function `AVG`. The discussion for `MAX` and `MIN` is similar, and omitted. Let x_i be the reading of node v_i and $\hat{x_i}$ the estimate of x_i. Consider two *highly correlated* neighboring nodes v_i and v_j, i.e. the absolute difference of their readings $|x_i - x_j|$ is no greater than ϵ, $0 \leq \epsilon < 1$. Assume that node v_i is the representative node of v_j in the representative routing tree. Then, x_i is an approximate value of x_j, and $|x_j - \hat{x_j}| = |x_j - x_i| \leq \epsilon$.

Let k be the number of nodes in the representative routing tree. Without loss of generality, assume the first k sensor nodes in the network are in the tree and node v_i represents n_i represented nodes including itself $1 \leq i \leq k$. Then, $\sum_{i=1}^{k} n_i = n$ following the definition of the representative routing tree. Let $y = \sum_{i=1}^{n} x_i/n$ be the exact average of the n sensor readings. Now, we use the representative routing tree to evaluate the `AVG` query to obtain an approximate solution of the query as follows.

If node v_i is a leaf node, it sends a message (N_i, W_i) to its parent, where $N_i = n_i$ is the number of nodes it represents and W_i is the sum of estimated values, i.e., either $W_i = n_i x_i$, or $W_i = x_i + \sum_{j=1}^{n_i - 1} \hat{x_j}$ and $\hat{x_j}$ is an estimate value of x_j, using the history data of node v_j that is stored at node v_i and the value of x_i itself. Note that the absolute difference between W_i and the sum of exact values of these nodes is no more than $(n_i - 1)\epsilon$, following the definition. If node v_i is an internal node, it sends a message (N_i, W_i) to its parent, where $N_i = \sum_{j=1}^{t_i} N_{i_j} + n_i$ and $W_i = \sum_{j=1}^{t_i} W_{i_j} + n_i x_i$, assuming that node v_i has t_i children and (N_{i_j}, W_{i_j}) is the message received from its jth child. Let (W, N) be the message received by the base station, then the estimated value is $\bar{y} = W/N = W/n$.

$$
\begin{aligned}
|y - \bar{y}| &= \left|\frac{x_1 + x_2 + \ldots + x_n}{n} - \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n}\right| \\
&\leq \frac{((n_1 - 1) + (n_2 - 1) + \ldots + (n_k - 1))\epsilon}{n} \\
&= \frac{(n-k)\epsilon}{n} = (1 - \frac{k}{n})\epsilon < \epsilon. \qquad \blacksquare
\end{aligned}
$$

V. Performance Study

We evaluate the performance of the proposed algorithms against that of an existing algorithm by conducting experimental simulations. The experimental results show that the proposed algorithms outperform the existing one significantly.

1) *Simulation environment* We assume that the monitored region is a $10 \times 10\ m^2$ square in which 1,000 homogeneous sensor nodes are deployed randomly, by the NS-2 simulator. We also assume that queries arrive one by one, and once a query arrives, it must be responded and evaluated by the system, using the established routing tree. For simplicity, we further assume that the length of the answer to each query is a unit-length. The energy consumption for evaluating a query using a representative routing tree as follows. Each node in the routing tree consumes t_e units of energy by transmitting one unit-length message to its parent, r_e units of energy by receiving one unit-length message from each of its children, and c_e units of energy by computing and processing the message from each node it represents. Each non-tree node that is represented by a tree node consumes t_e units of energy by transmitting its sensed data to its representative node (a tree node), assume that each non-tree node periodically sends its data every σ time units. An energy threshold δ is introduced to avoid that a representative routing tree may die quickly, since the lifetime of a routing tree is determined by the minimum ratio of the residual energy to the number of children at each node. The threshold δ will prevent those nodes that have lower residual energies to be chosen as the representative nodes in the routing tree. In other words, only those sensor nodes whose residual energies are no less than δ can be included in a representative routing tree.

To build a representative routing tree, the sensor nodes in the network are classified into either representative nodes or represented nodes, while the relationship between a representative node and its represented node is determined by the correlation threshold, using a given metric d. For convenience, we here assume that d is a function of overlapping sensing area between two nodes, where the sensing area by each sensor is a circle with radius r_s. The sensing *correlation percentage* between two neighboring nodes is the ratio of the overlapping sensing area to πr_s^2. Two neighboring nodes are *highly correlated* if the sensing correlation percentage between them is no less than a given correlation threshold θ, $0 < \theta < 1$. Each network topology with different problem size is generated using the NS-2 simulator. For each size of the network instance, the value shown in figures is the mean of 10 individual values obtained by running each algorithm on 10 randomly generated network topologies.

2) *Performance evaluation of various algorithms* We evaluate the performance of the proposed algorithms against that of an existing algorithm in terms of the network lifetime and the size of the representative routing tree, by varying correlation threshold and sensing range. In our simulation, the initial energy IE is 10^5 units and the energy threshold δ is 1% of IE. The transmission range is 2 and the energy consumption of per unit message at a node is as follows. The transmission energy consumption t_e is one unit, the reception energy consumption r_e and computing energy consumption c_e are 50% and 10% of t_e, respectively. The interval of communication (heart-beating) between a represented node and its representative node is 100 time units and the energy utility factor λ is 1.5.

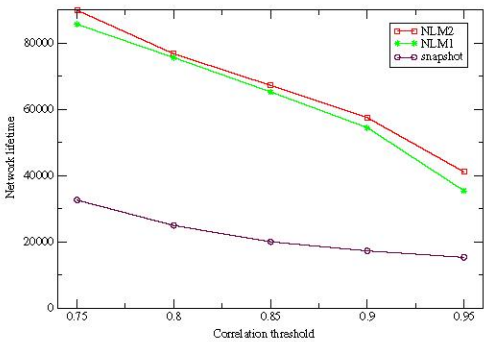

Fig. 1. Network lifetime evaluation with various correlation thresholds when $r_s = 1$.

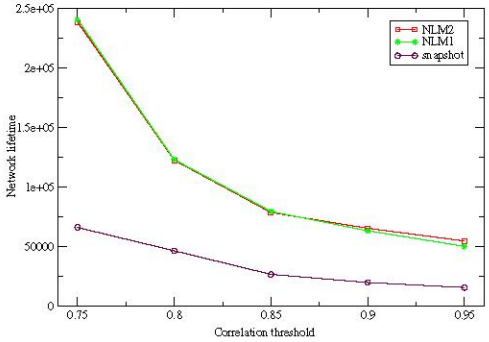

Fig. 2. Network lifetime evaluation with various correlation thresholds when $r_s = 1.5$.

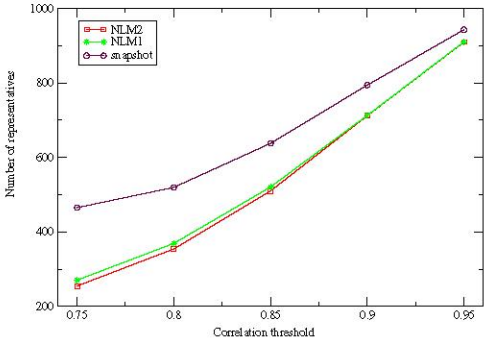

Fig. 3. Comparison of sizes of representative routing trees with various correlation thresholds when $r_s = 1$.

We first evaluate the network lifetime delivered by different algorithms with various data correlations when the sensing range r_s is 1, shown in Fig. 1. From this figure, we can see that algorithms `NLM1` and `NLM2` outperform algorithm `snapshot` significantly, when the correlation threshold varies from 75% to 95% with increment of 5%. The lifetime delivered by either algorithm `NLM1` or algorithm `NLM2` is nearly three times as long as that delivered by algorithm `snapshot` when correlation threshold is 75%. The lifetime delivered by algorithm `NLM1` or algorithm `NLM2` is still at least twice as long as that delivered by algorithm `snapshot` when the correlation threshold reaches 95%. In addition, we can see that with various correlation thresholds, the performance of algorithm `NLM2` is generally better than that of algorithm `NLM1`. This demonstrates that the exponent function based on the ratio of energy utilization models the energy consumption more precisely than the linear function of energy expiration. We then increase the sensing range r_s by 50% and evaluate the network lifetime delivered by each algorithm. Fig. 2 indicates that there is not any significant difference among the algorithms in terms of the performance, compared with the case where the sensing range r_s is 1, i.e. the performance of algorithms `NLM1` and `NLM2` are still constantly better than that of algorithm `snapshot`. We finally analyze the size of the representative routing trees, delivered by various algorithms through varying correlation thresholds. Fig. 3 implies that the tree delivered by algorithm `NLM1` or `NLM2` uses a fewer number of nodes than the one delivered by algorithm `snapshot`, when the correlation threshold is ranged from 75% to 95% with increment of 5%. The size of the representative routing tree delivered by algorithm `NLM2` is approximately 54% of that delivered by algorithms `snapshot` when the correlation threshold θ is 75%. From Fig. 3, we can see that the size difference among the representative routing trees delivered by different algorithms diminishes with the increase of the correlation threshold. When θ is 95%, although the size difference of representative routing trees delivered by algorithms `NLM2` and `snapshot` is around 33 only, the difference of network lifetime derived from them is as much as 25,800 time units, which can be seen from Fig. 1.

VI. Conclusions

In this paper, we considered the maximization problem of network lifetime for answering a sequence of unknown aggregate queries. We approached the problem by finding a series of nearly optimal representative routing trees for query evaluation. We first showed that the optimal representative routing tree problem is NP-Complete. We then presented heuristic algorithms for finding such representative routing trees. We finally conducted experiments by simulation to evaluate the performance of the proposed algorithms against an existing algorithm, in terms of network lifetime and average size of representative routing trees. The experimental results show that the proposed algorithms outperform the existing one significantly.

Acknowledgement. It is acknowledged that the work by the authors is fully funded by a research grant No:DP0449431 by Australian Research Council under its Discovery Schemes.

References

[1] R. Cristescu, B. Beferull-Lonzano, and M. Vetterli. On network correlated data gathering. *Proc. INFOCOM'04*, IEEE, 2004.

[2] A. Deshpande, C. Guestrin, S. Madden, J. M. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. *Proc. of VLDB*, 2004, 588–599.

[3] M. R. Garey and D. S. Johnson. *Computer and Intractability: A guide to the theory of NP-completeness*. W. H. Freeman, San Francisco, 1979.

[4] K. Kar, M. Kodialam, T. V. Lakshman, and L. Tassiulas. Routing for network capacity maximization in energy-constrained ad-hoc networks. *Proc. of INFOCOM'03.*, IEEE, 2003.

[5] Y. Koditis. Snapshot queries: towards data-centric sensor networks. *Proc. of ICDE'05*, IEEE, 2005.

[6] W. Liang and X. Guo. On-line multicasting for network capacity maximization in energy-constrained ad hoc networks. *IEEE Trans. Mobile Computing*, Vol. 5, No. 9, pp. 1215-1227,2006.

[7] S. Lindsey and C. S. Raghavendra. PEGASIS: Power-efficient gathering in sensor information systems. *Proc. Aerospace Conference*, IEEE, pp.1125–1130, 2002.

[8] S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong. TAG: a tiny aggregation service for ad hoc sensor networks. *ACM SIGOPS Operating Systems Review*, Vol. 36, pp.131–146, 2002.

[9] S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong. The design of an acquisitional query processor for sensor networks. *Proc. SIGMOD'03*, ACM, pp.491–502, 2003.

[10] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson. Wireless sensor networks for habitat monitoring. *Proc. of Int'l Workshop on Wireless Sensor Networks and Applications*, ACM, 2002.

[11] A. Meliou, D. Chu, C. Guestrin, and J. M. Hellerstein and W. Hong. Data gathering tours in sensor networks. *Proc.of IEEE IPSN'06*, 2006.

[12] P. von Rickenbach and R. Wattenhofer. Gathering correlated data in sensor networks. *Proc. 2nd DIALM-POMC*, ACM, Oct., 2004.

[13] F. Ye, G. Zhong, S. Lu, and L. Zhang. Peas: a robust energy conserving protocol for long-lived sensor networks. *Proc. of 23rd Int'l Conf. on Distributed Computing Systems*, IEEE, 2003.