An Efficient Node Selection Metric for In-network Process Deployment

Kenji Tei
Waseda University
3-4-1 Okubo, Shinjuku-ku
Tokyo, Japan
tei@aoni.waseda.jp

Yoshiaki Fukazawa
Waseda University
3-4-1 Okubo, Shinjuku-ku
Tokyo, Japan
fukazawa@waseda.jp

Shinichi Honiden
National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku
Tokyo, Japan
honiden@nii.ac.jp

ABSTRACT

In-network processing is a powerful technique for reducing network traffic in an ad hoc network where network efficiency is a critical issue. When an in-network process collects data from multiple data sources, the node hosting the in-network process should be carefully selected to reduce network traffic. Existing metrics used to select the host node are unsatisfactory in this case, because they do not consider differences in the amount of data provided by each data source. In this paper, we propose a node selection metric called COLOR to solve this problem. COLOR value is derived from locations of data sources and the amount of data provided by them so that a data source that provides more data than the others has a stronger effect. Moreover, the communication overheads associated with COLOR are small, because parameters involved by COLOR can be collected during a data retrieval phase, which generally occurs in in-network processing. Simulation results show that data retrieval using COLOR produces less network traffic than that retrieved using existing metrics in environments where placements of data sources and the amount of data are nonuniform.

Categories and Subject Descriptors

C.2.4 [**Computer Systems Organization**]: Computer-Communication NetworksDistributed Systems[Distributed Application]; D.2.8 [**Software Engineering**]: Metrics—*performance measures*

General Terms

Management

Keywords

Mobile Ad hoc Network, In-network Processing

WICON '08 Hawaii USA

1. INTRODUCTION

A wireless ad hoc network is a decentralized network consisting of autonomous nodes communicating with each other using wireless links. One potential application of such a network is for carrying out surveys in disaster areas. In a disaster area, data from data sources (i.e. sensor networks or RF-ID repositories) providing data about the disaster area are extremely useful, but may be difficult to retrieve because communication infrastructure may be damaged by the disaster and the data sources may be isolated from the Internet. Ad hoc networks, consisting of smart phones or PDAs, can be used as alternative networks. An observer sends queries to nodes near the data sources via the network. Then, the nodes reply by sending data collected from the data sources using a short-range wireless communication device such as ZigBee or an RF-ID reader. Finally, the observer analyzes the retrieved data and produces a report on the area.

In an ad hoc network, network traffic caused by data retrieval should be small, because the battery power and bandwidth of mobile nodes constituting the network are limited. In-network processing[4, 6, 10, 11] is a widely used method of reducing network traffic by executing data aggregation or data fusion in the network. More concretely, a software processes called processing element (PE) is deployed at a node near data sources. Then a PE collects data from nodes that can directly access the data sources called accessible nodes. The PE periodically executes data aggregation or data fusion, and finally sends the result to the observer. Compared with the case where all raw data is sent independently to an observer, in-network aggregation produces low network traffic, since only data processing results, which generally contain less data than the raw data, are sent to the observer. Therefore, network traffic can be reduced.

To apply in-network processing to data retrieval in ad hoc networks, a node hosting a PE should be carefully selected to sufficiently reduce network traffic. Using in-network processing, network traffic between an observer and a PE-hosting node can be reduced. However, network traffic between a node hosting the PE and accessible nodes cannot be reduced, because data transferred between them is raw data. Therefore, a PE should be deployed on a node to ensure that the network traffic from accessible nodes is small.

To select a suitable node, a node selection metric fulfilling two requirements is required. The first requirement is a highly accurate estimation of network traffic. The metric should accurately estimate actual network traffic caused by

Digital Object Identifier: 10.4108/ICST.WICON2008.4851
http://dx.doi.org/10.4108/ICST.WICON2008.4851

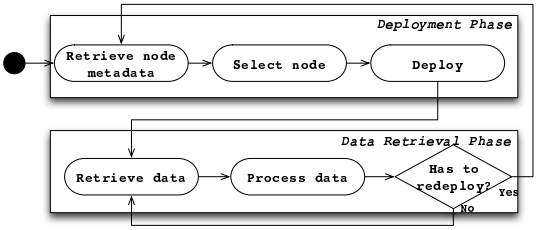

Figure 1: Flowchart of PE

communication between a PE-hosting node and accessible nodes. The second requirement is that the metric should have low communication overheads for retrieving input parameters for the estimation.

Existing node selection metrics are node-centric. They focus on network topology[2, 3, 13] or the locations of nodes[9, 1, 11]. However, they may not select a suitable node in a realistic environment. They assume that all data transferred in a network is of the same size, or that all communication links between any two nodes have same traffic, neither of which is a realistic assumption. In a realistic environment, data may be more heterogeneous. The size of data sent from accessible nodes depends on the data sources, and network traffic between any two nodes depends on the amount of data transferred and the communication route of the data. Therefore, existing metrics cannot estimate actual network traffic in an environment with data heterogeneity.

In this paper, we propose a data-centric node selection metric called COLOR (Cost Of LOcation for Relocation). COLOR focuses on data sources. It involves the location of each data source and the amount of data it provides, as parameters. Then, it estimates network traffic from these parameters and the location of the node, and outputs a value representing the traffic, taking into account the data heterogeneity. Moreover, the parameters of COLOR can be collected with lower cost than the parameters of existing node-centric metrics.

The rest of this paper is organized as follows. In section 2 we describe the node selection problem in detail, and in section 3 we describe existing metrics. In section 4 we describe COLOR in detail. In section 5 we discuss the overheads of COLOR and in section 6 we report the results of its evaluation. Finally, we conclude the paper in section 7.

2. PRELIMINARY

Here we provide a brief overview of PE activities to clarify the target of node selection and its formulation.

2.1 Brief overview of PE activities

Figure 1 illustrates a typical flowchart of PE activities shown in [1, 11]. PE activities are classified into a data retrieval phase and a deployment phase.

In the data retrieval phase, a PE broadcasts queries to retrieve raw data. A node receiving the query tries to find data sources that it can access. If it finds at least one data source (then the node is an accessible node), it accesses the data sources to obtain raw data using short-range wireless communication and replies to the node hosting the PE by sending the raw data. The PE aggregates or fuses the retrieved raw data. The data retrieval phase is periodically executed to adapt to changes in the network topology.

A PE can be redeployed at another node to maintain low communication traffic produced by raw data retrieval. Even if a PE is initially deployed at a suitable node, network traffic between the node and accessible nodes may increase according to changes in the network topology and changes in the amount of data provided by each data source. It is periodically decided whether a PE should be redeployed or not. Conditions affecting this decision are the passage of a certain period of time[8], the number of hops from other nodes[4], and the location and speed of the node[11]. These conditions are adjusted with the aim of eliminating needless redeployment, since redeployment itself causes network traffic. In this paper, we adopt these existing conditions and do not consider the conditions in detail to focus on the node selection problem described below.

When a PE is to be redeployed, the deployment phase is executed. In the deployment phase, first the PE collects data on nodes, second the PE evaluates each node in accordance with a node selection metric and the data on the nodes, Then, the PE selects the node with the best value derived from the metric as the next PE-hosting node. Finally, the PE stops its activities, sends its program code and fused data to the next node, and restarts its activities on the node.

To reduce network traffic, the selection of the PE-hosting node is important. Network traffic caused by the PE activities constitutes mainly raw data retrieval involving communication between a PE-hosting node and accessible nodes. Therefore, we focus on the node selection problem that will strongly affect network traffic in this paper.

2.2 Formulation of data retrieval

In this section, we discuss the node selection problem in detail. In a two-dimensional field, we consider q nodes ($N = \{n_1, n_2, ..., n_q\}$) with locations $L_N = \{\vec{l_{n_1}}, \vec{l_{n_2}}, ..., \vec{l_{n_q}}\}$. Since each node can move, $\{\vec{l_{n_1}}, \vec{l_{n_2}}, ..., \vec{l_{n_q}}\}$ may change over time. The location of each node is determined using a location estimation method, such as a global positioning service (GPS).

In the field, there are p data sources (represented by $S = \{s_1, s_2, ..., s_p\}$), whose corresponding locations are $L_S = \{\vec{l_{s_1}}, \vec{l_{s_2}}, ..., \vec{l_{s_p}}\}$. Each data source does not need to determine its own location. The amounts of data collected from the data sources are $A_S = \{a_{s_1}, a_{s_2}, ...a_{s_p}\}$. Nodes located within distance d_i from s_i are the nodes accessible to s_i, and can collect a_{s_i}-sized data from s_i.

Figure 2 illustrates an overview of data retrieval by a PE. Let n_c ($n_c \in N$) be the current PE-hosting node. In figure 2, there are 12 nodes $N = \{n_1, n_2, ..., n_{12}\}$ and 5 data sources $S = \{s_1, s_2, ..., s_5\}$, where $n_c = n_4$. Dotted circles centered at each data source represent the communication ranges of the data source. In figure 2, there are a total of 9 accessible nodes $N_{s_1} = \{n_1, n_2, n_3\}$, $N_{s_2} = \{n_6, n_7, n_8\}$, $N_{s_3} = \{n_{10}, n_{11}\}$, and $N_{s_5} = \{n_{12}\}$. Note that the PE cannot collect data from s_4 since it has no accessible node, and it also cannot collect data from s_5 since although it has an accessible node (n_{12}), no communication route between s_c and n_{12} exists. Therefore, in the network illustrated in figure 2, the PE can collect data on only the 3 data sources (s_1, s_2, s_3) from the 8 accessible nodes. To reduce network traffic, a suitable n_c should be selected.

Digital Object Identifier: 10.4108/ICST.WICON2008.4851
http://dx.doi.org/10.4108/ICST.WICON2008.4851

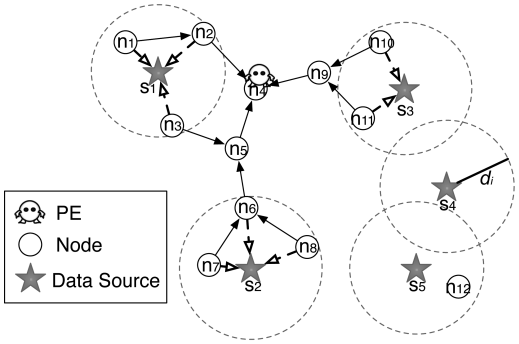

Figure 2: Overview of data retrieval

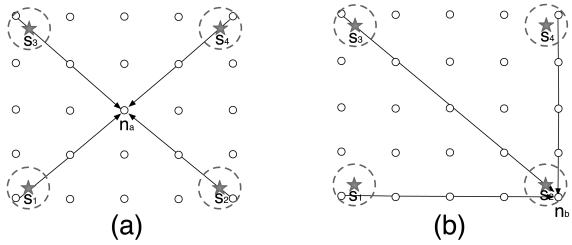

Figure 3: Simple environment for data retrieval

3. RELATED WORKS

Existing node selection metrics focus on network topology[2, 3, 13] or the locations of nodes[1, 9, 11].

Topology-based metrics are node selection metrics based on network topology. DCA[3], the metric proposed in [2], and TASC[13] are examples of such metrics. DCA focuses on the number of one-hop neighbor nodes. DCA selects the node that has the most neighbors as n_c. DCA can select a node that has the least possibility to isolate from the network, but it cannot optimize network traffic. The metric proposed in [2] focuses on the number of communication hops. The metric selects a node that can communicate with any node within a designated number of hops as n_c, but the optimal number of hops is not described in [2]. TASC focuses on network topology. TASC selects the node that is most frequently used in communication routes between any two nodes, as n_c. The derivation of this node in accordance with TASC is as follows.

1. Given two nodes, the weights of all nodes along the shortest route between them are increased by one.
2. Carry out step 1 for all any two nodes.
3. Finally, select the node having the biggest weight as n_c.

Moreover, TASC considers the distance between nodes. In step 1 described above, the weight can be increased not by one, but by w as described below. If n_k lies in the route from n_i to n_j between n_a and n_b, then the weight increase of n_k is given by equation (1), where $distance_{a,k}$ and $distance_{k,b}$ are the distances between n_a and n_k, and between n_k and n_b, respectively, and $distance_{ij}$ is the distance of the whole route from n_i to n_j.

$$w = \frac{distance_{a,k} + distance_{k,b}}{distance_{ij}} \quad (1)$$

Using TASC, communication routes between n_c and other nodes can be optimized.

Location-based metrics are node selection metrics based on node location. GRID[9], GeoBee[11], and EnviroTrack[1] are examples of such metrics. In GRID and GEOBEE, the node nearest the barycenter of a designated field is selected as n_c. In EnviroTrack, the node nearest the barycenter of nodes that communicate with the last n_c, is selected as the next n_c.

However, these existing metrics may not select a suitable node in environments with data heterogeneity. Consider the case illustrated in figure 3. Twenty-five nodes are placed on a 5×5 grid on a square field with lower-left and top-right coordinates of $(0, 0)$ and $(200, 200)$, respectively. Each node can communicate with neighboring nodes. Moreover, 4 data sources $s_1, s_2, s_3, ands_4$ are placed at $(10, 10)$, $(190, 10)$, $(10, 190)$, and $(190, 190)$, respectively, and their communication ranges are $d_1 = d_2 = d_3 = d_4 = 15m$.

According to TASC, which is based on network topology, n_a, which is the central node of the network topology, is selected as n_c, According to EnviroTrack, which is based on node location, n_a, which is the node nearest the barycenter of all the nodes, is selected as n_c. However, when s_2, which is the lower-right data source in figure 3, provides more data than the other data sources, n_a is not the best node. Let $A_S = (d, 5d, d, d)$, and the network traffic be defined as the product of the data size and the number of communication hops. The network traffic when $n_c = n_a$ is $16d$, whereas that when $n_c = n_b$ is $12d$. Therefore, network traffic when $n_c = n_b$ is about 25% smaller than that when $n_c = n_a$. Therefore, existing metrics are unsuitable in environments with data heterogeneity, where each data source provides a nonuniform amount of data.

4. COLOR

In this section, we describe our proposed COLOR node selection metric in detail.

4.1 Detail of COLOR

COLOR is a data-centric node selection metric that focuses on data sources to select n_c. It uses the location of each data source and the amount of data provided by each data source as parameters. The COLOR value of n_i is affected by s_j and is proportional to both the distance between n_i and s_j and the amount of data provided by s_j. When the distance between n_i and s_j increases, the COLOR value increases since the expected number of hops involved in the data retrieval phase increases. When the amount of data provided by s_j increases, the COLOR value increases since the amount of data transferred in the network increases. Thus, the COLOR value depends on weighted distances derived from the above two factors.

The derivation of the COLOR value is as follows. For each data source from which a PE can collect data, a weighted distance is derived. The sum of the weighted distances represents the COLOR value. More specifically, a function $F_{COLOR}(\vec{l})$ to derive the COLOR value of a node located at $\vec{l}$ is given by equation (2), where the set of data sources from

Digital Object Identifier: 10.4108/ICST.WICON2008.4851
http://dx.doi.org/10.4108/ICST.WICON2008.4851

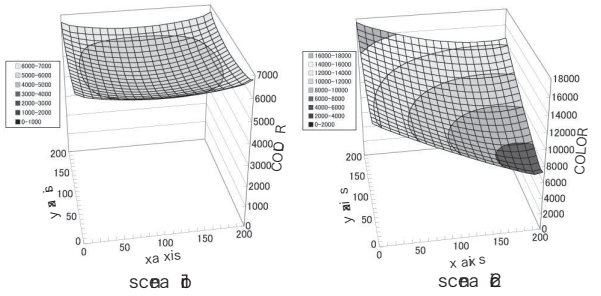

Figure 4: Distribution of COLOR in the examples

which a PE can collect data is $S' = \{s_1, s_2, ..., s_{p'}\}(S' \subseteq S)$, and their locations and the amount of data are $L_{S'} = \{\vec{l_{s'_1}}, \vec{l_{s'_2}}, ..., \vec{l_{s'_{p'}}}\}$ and $A_{S'} = \{a_{s_1}, a_{s_2}, ..., a_{s_{p'}}\}$, respectively.

$$F_{COLOR}(\vec{l}) = \sum_{s_i \in S'} (a_{s_i} \times distance(\vec{l}, \vec{l_{s_i}})) \quad (2)$$

In COLOR, the node having the smallest value derived from equation (2) is selected as n_c.

Figure 4 illustrates the distribution of the COLOR value in two scenarios. In scenario 1, each data source provides 10kB of data. Scenario 2 is the same as scenario 1 except that s_2 provides 50kB of data. In scenario 1, COLOR selects the node nearest (100,100) as n_c, which is the barycenter of the 4 data sources, since the data sources provide the same amount of data. Whereas in scenario 2, COLOR selects the node nearest (190,10) as n_c, which is nearer s_2 than n_c in scenario 1, since s_2 provides more data than the other data sources. In COLOR, n_b is selected as n_c, which is the same result as that shown in section 2.2.

When it is decided that a PE is to be redeployed, the PE sends queries to collect the current location of each node and derives its COLOR values using $F_{COLOR}(\vec{l})$ and the collected location. Consequently, the node having the smallest COLOR value is then selected as n_c.

4.2 Collecting COLOR parameters

Equation (2) involves the location of each data source from which a PE can collect data ($L_{S'}$), and the amount of data provided by it ($A_{S'}$). $L_{S'}$ and $A_{S'}$ can be dynamically determined in the data retrieval phase. The location of a data source is approximated by the barycenter of the nodes accessible to the data source, and the amount of data provided by a data source is approximated by the total amount of data sent from the nodes accessible to the data source.

$L_{S'}$ and $A_{S'}$ are determined as follows. In the data retrieval phase, the nodes accessible to s_i send raw data provided by it to the PE. At this time, the node contains its own location in a message also containing the raw data. The location of s_i ($\vec{l_{s_i}}$) is determined from the locations of nodes accessible to s_i. Let the locations of nodes accessible to s_i be $L_{N_{s_i}} = (\vec{l_{n'_1}}, \vec{l_{n'_2}}, ..., \vec{l_{n'_{p'}}})$, where $\vec{l_{s_i}}$ is given by equation (3).

$$\vec{l_{s_i}} = \frac{\sum_{\vec{l} \in L_{N_{s_i}}} (\vec{l})}{p'} \quad (3)$$

Moreover, let the amount of data sent from the nodes accessible to s_i be $A_{N_{s_i}} = (a_{n'_1}, a_{n'_2}, ..., a_{n'_{p'}})$, where a_{s_i} is given by equation (4).

$$a_{s_i} = \sum_{a \in A_{N_{s_i}}} (a) \quad (4)$$

$L_{S'}$ and $A_{S'}$ are determined by evaluating equations (3) and (4) for each data source, respectively. The determined values of $L_{S'}$ and $A_{S'}$ are then used in equation (2).

5. DISCUSSION

In this section, we discuss the communication and computational overheads of COLOR compared with TASC, a representative topology-based metric, and with EnviroTrack, a representative location-based metric.

5.1 Communication overheads

Node selection metrics involve input data on nodes and parameters. However, collection of the input data and parameters increases network traffic, since it involves network communications. To reduce the total network traffic, the parameters should be collected with a low communication overhead.

We first discuss the communication overhead for collecting the input data. Input data is collected in the deployment phase. The input data of both EnviroTrack and COLOR constitutes the locations of nodes. In both cases, a PE sends queries to nodes and retrieves their locations. The input data of TASC constitutes the network topology. In this case, a PE sends queries to nodes and retrieves the parts of the network topology contained in each node. From the viewpoint of data size, the data size for node location would be smaller than that of part of the network topology. When the location of a node is represented by two-dimensional coordinates, its data size would be at most a few dozen bytes. When the network topology is represented by a set of pairs of IP addresses, even if the network consists of only two nodes, its data size would be about 16 bytes. Furthermore, the data size increases at a rate proportional to the square of the number of nodes. Therefore, the communication overhead for collecting the input data of COLOR is less than or equal to that of the other existing metrics.

We now discuss the communication overhead for collecting the parameters. The parameters of COLOR are L_S and A_S, which, as described in Section 4.2, can be collected in the data retrieval phase. They are estimated from the locations of accessible nodes and the amount of collected data, respectively. Since the location of an accessible node is contained with raw data in a message, and since the amount of collected data is derived from the raw data, no additional queries are necessary to collect the parameters. The parameter of EnviroTrack is also the location of accessible nodes, which can be collected in the data retrieval phase, as with COLOR. Therefore, the communication overhead for collecting parameters of COLOR is the same as that of EnviroTrack. On the other hand, the parameter of TASC is part of the network topology. The communication overhead for estimating the network topology is huge, since every node must periodically send a hello message to its one-hop neighbor nodes. In [12], it is shown that the communication overhead for estimating a network topology is nearly half the total communication overhead of data retrieval when an observer retrieves 10kB data from 16 data sources every minute via

Digital Object Identifier: 10.4108/ICST.WICON2008.4851
http://dx.doi.org/10.4108/ICST.WICON2008.4851

an ad hoc network consisting of 16 nodes in a 1km square field, and every node sends a hello message to its neighbors every 30 seconds. For the OLSR protocol[7], which is a routing protocol also involving a network topology, it is recommended that the frequency of topology estimation is set to 2 seconds. This would require 15 times the communication overhead than that in the case of [12]. Therefore, the communication overhead for collecting the parameters of COLOR is smaller than that of TASC. As a whole, the total communication overhead of COLOR is the same or smaller than that of EnviroTrack and TASC.

5.2 Computational overhead

Next, we compare the computational overhead of COLOR with that of EnviroTrack and TASC. The computational overhead for the selection of n_c in the deployment phase increases the delay of redeployment. Therefore, it should be small.

The computational overhead of COLOR consists of the overhead for calculating equation (2) and is expressed as pqC, where p is the number of data sources that can be accessed via the network, q is the number of nodes, and C is the computational cost of calculating the distance between two nodes. To calculate equation (2), locations of data sources and the amount of data provided by them must be estimated, but the computational overhead for estimating data sources does not affect the computational overhead in the deployment phase, since these estimations are executed in the data retrieval phase, as discussed in section 4.2.

The computational overhead of EnviroTrack is expressed as $q'qC$, where q' is the number of accessible nodes. Since the number of accessible nodes is greater than or equal to that of the data sources that can be accessed via the network, $q' \geq p$. Therefore, the computational overhead of COLOR is smaller than or the same to that of EnviroTrack.

The computational overhead of TASC, which calculates equation (1) for all pairs of nodes, is expressed as $q^2(3C)$. Since the number of nodes is greater than the number of data sources in many cases, $q > p$. Therefore, the computational overhead of COLOR is smaller than that of TASC.

6. EVALUATION

In this section, we evaluate the COLOR metric. We first evaluate accuracy of COLOR by comparison with simulation results. Second, we evaluate the network traffic in the data retrieval phase using COLOR, in comparison with that using EnviroTrack and TASC. Finally, we evaluate the total network traffic in dynamic environments.

6.1 Simulation setting

Simulations are performed on a simulator implemented upon the *SWANS* framework[14]. The simulation setup is as follows. As shown in scenario 2 in section 4.1, 4 data sources $(s1, s2, s3, s4)$ are placed in 200m square field with lower-left and top-right coordinates of$(0, 0)$ and $(200, 200)$, respectively. The locations of the data sources are $\vec{l_{s_1}} = (10, 10)$, $\vec{l_{s_2}} = (190, 10)$, $\vec{l_{s_3}} = (10, 190)$, and $\vec{l_{s_4}} = (190, 190)$, and the amounts of data provided by the data sources are $a_{s_1} = a_{s_3} = a_{s_4} = 10$kB and $a_{s_2} = 50$kB. Nodes within 15m from a data source are accessible to the data source. Each node is equipped with an IEEE 802.11 wireless communication device, and can communicate with nodes within about 70m.

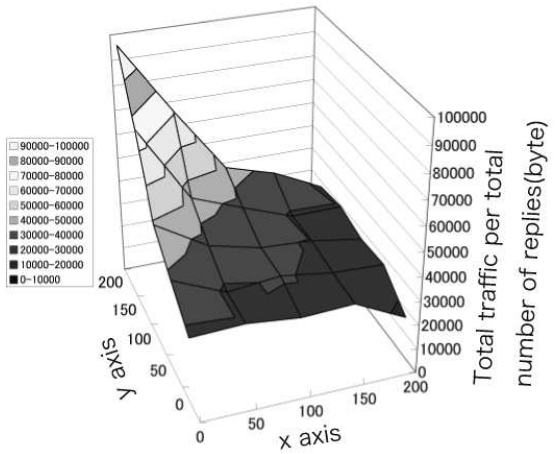

Figure 5: Network traffic per the total number of replies (for different values of n_c)

Table 1: Comparison of simulation results using COLOR

$\vec{l_{n_c}}$	$F_{COLOR}(\vec{l_{n_c}})$	simulation results
$(100, 100)$	10748	30880
$(200, 0)$	7012	16779
ratio	-34.8%	-45.7%

A PE sends a query to retrieve data every 30 seconds for 30 minutes. An accessible node that receives the query retrieves data from its accessible data source, and replies the retrieved data to the PE. We measure the total number of replies and total network traffic.

6.2 Accuracy of COLOR

In this section, we evaluate accuracy of COLOR by comparison with simulation results. We compared the network traffic estimated by COLOR for the environment shown in figure 3 with the network traffic derived by a simulation. In the simulation, 25 nodes $(n_1, n_2, ..., n_{25})$ are placed on a 5×5 grid (as for scenario 2 in figure 3). For each n_i, 50 simulations are executed in which $n_c = n_i$. To compare only network traffic caused by data retrieval, all nodes are fixed, and the PE is not redeployed. Figure 5 illustrates the network traffic per the total number of replies, for each L_i with i from 0 to 25.

The simulation results illustrated in figure 5 show a similar trend to the estimation by COLOR illustrated in figure 4. In both cases, the network traffic per the total number of replies is the minimum when n_c is located at $(200, 0)$ which is the nearest node to s_2, and it increases with the distance between n_c and s_2. The rates of increase in figures 5 and 4 are different. Table 1 shows the estimation by COLOR and the simulation results when $n_c = n_a$ or $n_c = n_b$ ($l_{n_a} = (100, 100)$ and $n_b = (200, 0)$). In the case of the estimation by COLOR, the result when $n_c = n_a$ is 35.8% smaller than that when $n_c = n_b$, whereas in the case of the simulation results, the result when $n_c = n_a$ is 45.7% smaller than that when $n_c = n_b$. This is because in equation (2) it is assumed that the COLOR value increases accord-

Digital Object Identifier: 10.4108/ICST.WICON2008.4851
http://dx.doi.org/10.4108/ICST.WICON2008.4851

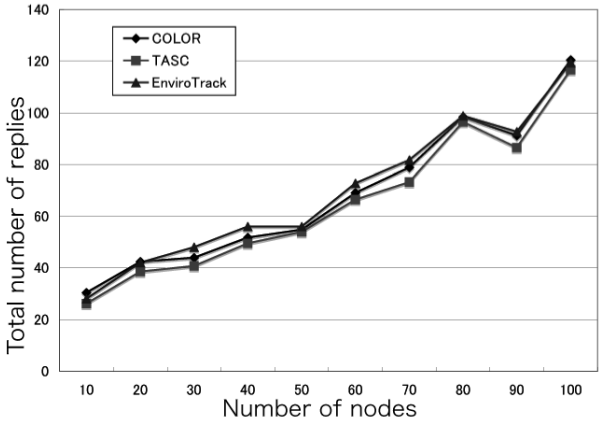

Figure 6: The total number of replies (for different values of density)

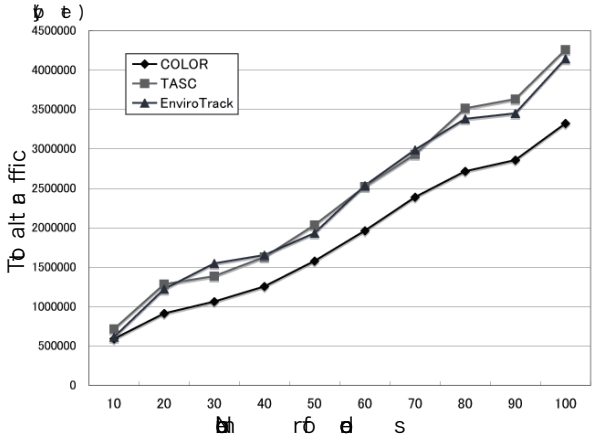

Figure 7: The total amount of transferred data (for different values of density)

ing to the distance between the PE-hosting node and the data source, but actual network traffic increases more due to packet losses and retransmissions. Therefore, although equation (2) can be used to estimate the trend of actual network traffic, its accuracy of estimation is not high.

However, accuracy is not a major issue in the selection of n_c. The only requirement of a node selection metric is to determine the optimal node for n_c, not to determine the ratio accurately. In this situation, COLOR can determine n_b which is the optimal node in this situation, as n_c, in contrast to other metrics such as EnviroTrack and TASC, as discussed in section 3. Therefore, COLOR is a suitable node selection metric in this situation.

6.3 Effectiveness of data retrieval in static environments

In this section, we evaluate COLOR from the point of efficiency of data retrieval, in comparison with EnviroTrack and TASC. In this simulation, nodes are placed randomly and are fixed. The total number of nodes is varied from 10 to 100. Figure 7 illustrates the total network traffic, figure 6 illustrates the total number of replies, and figure 8 illustrates the total network traffic per the total replies. Table 2 shows the results for COLOR relative to those of TASC and EnviroTrack.

We compared the metrics from the viewpoint of the total number of replies. Figure 6 shows that the total number of replies of COLOR is about 5.47%, on average, larger than that of TASC. COLOR and EnviroTrack select n_c based on the locations of data sources whose data can be retrieved via the network, but TASC selects n_c based on node topology. Therefore, TASC might select an n_c in which the PE can retrieve few raw data from data sources.

Second, we compared the metrics from the viewpoint of the total network traffic. Figure 7 shows that the total network traffic of COLOR, EnviroTrack, and TASC are about the same when the number of nodes is 10. In this case, the reduction of the network traffic using COLOR is small, since there are only a few accessible nodes. However, with increasing number of nodes, data retrieval using COLOR produces less network traffic than the other metrics. When the number of nodes is from 40 to 100, the total network traffic of COLOR is about 28.7% and 26.2% smaller than that of TASC and EnviroTrack, respectively. When there are many nodes, the reduction of the network traffic using COLOR becomes large since there are many accessible nodes. TASC considers only network topology. If all nodes send the same amount of data, TASC might be able to select an optimal n_c. However, in many practical cases, each node will send a different amount of data, since only accessible nodes send data whose size will be different. In addition, EnviroTrack only considers the locations of node sources. In many practical cases, each data source provides a different amount of data, which might vary with time. Therefore, COLOR, which considers both the locations of and the amounts of data provided by data sources, produces less network traffic than the other metrics.

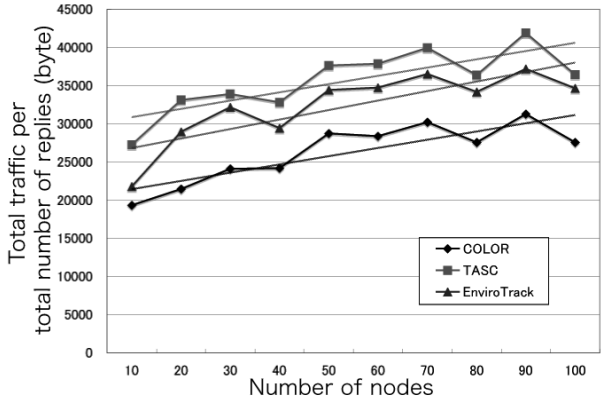

Figure 8: Network efficiency in the static environment

Finally, we compared the metrics from the viewpoint of network efficiency. We adopted the total network traffic per the total number of replies as a metric for the comparison. The straight lines in figure 8 are approximate lines of best fit

Digital Object Identifier: 10.4108/ICST.WICON2008.4851
http://dx.doi.org/10.4108/ICST.WICON2008.4851

Table 2: COLOR results as relative values in the static environment(%)

# of nodes	10	20	30	40	50	60	70	80	90	100	average
Traffic to TASC	-21.1	-40.5	-30.4	-29.8	-29.0	-28.4	-22.7	-29.5	-27.1	-28.1	-28.7
to EnviroTrack	-4.1	-33.8	-45.7	-31.7	-22.5	-29.1	-25.2	-24.5	-20.8	-24.7	-26.2
Inforamtion to TASC	14.04	8.98	7.34	4.28	1.52	3.86	7.21	1.95	5.13	3.11	5.47
to EnviroTrack	7.69	0.87	-9.13	-8.27	-2.23	-5.38	-3.61	-0.44	-1.56	0.81	-2.12
Efficiency to TASC	-41	-54.4	-40.7	-35.6	-31	-33.5	-32.2	-32.1	-34	-32.2	-36.7
to EnviroTrack	-12.7	-35	-33.5	-21.6	-19.8	-22.5	-20.8	-24	-18.9	-25.8	-23.5

for the results, drawn by the least-squares method. Figure 8 shows that COLOR is on average 36.7% and 23.5% more efficient than TASC and EnviroTrack, respectively, and particularly efficient in environments where data sources provide nonuniform amounts of data.

6.4 Effectiveness of data retrieval in dynamic environments

In this section, we evaluate COLOR from the viewpoint of the efficiency of data retrieval in dynamic environments including redeployment, in comparison with EnviroTrack and TASC. In this simulation, nodes are initially placed randomly and move in accordance with a random walk model[5]. A random direction is chosen for each node which moves $0-10$m every 10 seconds. The number of nodes is varied from 10 to 100. To compare the effectiveness of the node selection metrics, a simple condition for redeployment proposed in figure [8] is adopted in all cases that depends on the passage of a fixed period of time. In these simulations, a PE is redeployed every 30 seconds. Additionally, in the case of TASC, each node sends a hello message to its neighbors to capture part of the network topology. Figure 9 illustrates the network efficiency in dynamic environments.

The result shows that COLOR is relatively more efficient in a static environment. When the number of nodes is 100, COLOR is 77.6% more efficient than TASC, whereas it is only 32.2% more efficient in the static environment shown in section 6.3. This is because TASC involves parameters that produce a large amount of traffic in a dynamic environment. With increasing number of nodes, the network traffic caused by the topology estimation increases, which reduces the network efficiency of TASC. However, COLOR involves parameters that produce less traffic even in a dynamic environment. As discussed in section 5.1, the parameters of COLOR can be retrieved in the data retrieval phase with a low communication overhead. Therefore, improvement of efficiency of COLOR compared with TASC in a dynamic environment is larger than that in a static environment since parameters of COLOR can be collected with a low communication overhead.

When the number of nodes is 100, COLOR is 32.5% more efficient than EnviroTrack, whereas it is 25.8% more efficient in the static environment shown in section 6.3. This is because EnviroTrack may select an n_c that is less suitable than the previous n_c since it cannot estimate network traffic. However, COLOR can select a more suitable n_c than the previous n_c since it can estimate network traffic in adapting to network changes. Consequently, COLOR is more suitable in dynamic environments than in static environments.

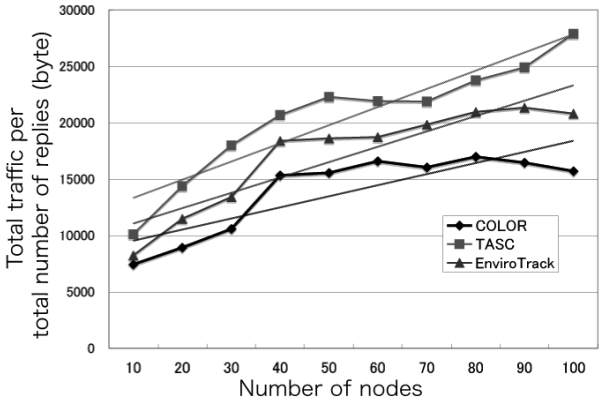

Figure 9: Network efficiency in the dynamic environment

7. CONCLUSION

In this paper, we propose a node selection metric named COLOR for efficient in-network processing. COLOR is a data source-centric metric and can estimate network traffic more accurately than existing node-centric metrics in environments with data heterogeneity in which data sources provide nonuniform amounts of data. Moreover, COLOR involves parameters that can be collected with a low communication overhead. Therefore, COLOR is particularly suitable for dynamic environments in which the network state changes dynamically, compared with other metrics that involve parameters that can be collected with a high communication overhead. Using COLOR for the redeployment of a PE, the PE can maintain low network traffic for retrieving data from data sources, even in a dynamic environment.

To further improve the network efficiency, we should consider not only a node selection metric for redeployment but also a combination of the metric and a condition to decide the redeployment. In the deployment phase, a large amount of data, containing the PE program and fused data, is transferred in the network. From the viewpoint of network traffic caused by the redeployment, the frequency of redeployment should be small. To reduce the total network traffic, the redeployment should be carefully determined on the basis of the network traffic generated by data retrieval after the redeployment and network traffic caused by the redeployment itself. We focus on this problem in future works.

8. REFERENCES

[1] T. Abdelzaher and et al. Envirotrack: Towards an environmental computing paradigm for distributed sensor networks. In *Proceedings of the 24th International Conference on Distributed Computing Systems (ICDCS)*, pages 582–589, 2004.

[2] A. D. Amis and et al. Max-min d-cluster formation in

Digital Object Identifier: 10.4108/ICST.WICON2008.4851
http://dx.doi.org/10.4108/ICST.WICON2008.4851

wireless ad hoc networks. In *Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, pages 32–41, 2000.

[3] S. Basagni. Distributed clustering for ad hoc networks. In *Proceedings of the 1999 International Symposium on Parallel Architectures, Algorithms and Networks (ISPAN)*, pages 310–315, 1999.

[4] B. Blum and et al. An entity maintenance and connection service for sensor networks. In *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services (MobiSys)*, pages 201–214, 2003.

[5] T. Camp, J. Boleng, and V. Davies. A survey of mobility models for ad-hoc network research. *Wireless Communications and Mobile Computing*, 2(5):483–502, 2002.

[6] G. Chen and D. Kotz. Policy-driven data dissemination for context-aware applications. In *Proceedings of the Third IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 283–289, 2005.

[7] T. Clausen and P. Jacquet. Optimized Link State Routing Protocol (OLSR). RFC 3626 (Experimental), Oct. 2003.

[8] W. R. Heinzelman and et al. Energy-efficient communication protocol for wireless microsensor networks. In *Proceedings of the 33rd International Conference on System Sciences-Volume 8 (HICSS)*, page 8020, 2000.

[9] W.-H. Liao and et al. Grid: A fully location-aware routing protocol for mobile ad hoc networks. *Telecommunication Systems*, 18(1-3):37–60, 2001.

[10] S. Madden and et al. Tag: a tiny aggregation service for ad-hoc sensor networks. In *Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI)*, pages 131–146, 2002.

[11] K. Tei and et al. Using mobile agent for location-specific data retrieval in manet. In *Proceedings of the 2005 IFIP International Conference on Intelligence in Communication Systems (INTELLCOMM)*, pages 157–168, 2005.

[12] K. Tei and et al. Adaptive geographically bound mobile agents. In *Proceedings of the 2nd International Conference on Mobile Ad-hoc and Sensor Networks (MSN)*, pages 353–364, 2006.

[13] R. Virrankoski and A. Savvides. Tasc: Topology adaptive spatial clustering for sensor networks. In *Proceedings of the 2nd IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS)*, pages 10–18, 2005.

[14] X. Zhang and G. F. Riley. Scalability of an ad hoc on-demand routing protocol in very large-scale mobile wireless networks. *Simulation*, 82(2):131–142, 2006.

Digital Object Identifier: 10.4108/ICST.WICON2008.4851
http://dx.doi.org/10.4108/ICST.WICON2008.4851