2009－UBI－21 (13)
2009／3／4

センサネットワークのためのトポロジーの変化を考慮したデータ集約方式

柳沢 豊　櫻井 保志　岸野 泰恵　前川 卓也　亀井 剛次　岡留 剛

日本電信電話株式会社，NTT コミュニケーション科学基礎研究所
〒 619–0237 京都府相楽郡精華町光台 2-4

概要

空間的な相関性のあるデータのモニタリングを行うセンサネットワークにおいて，ネットワークトポロジーの変化が生じても，通信コストを最小限に抑えることができるデータ集約方式を提案する．提案手法では，ネットワークを複数のクラスタに分割してクラスタごとに相関性のあるデータを統合する方法を用いる．このとき発生する，統合されたデータに符号を割り当てる際に大きな記憶領域を必要とする問題について，サイズの上限が設定されたキャッシュを用いることによる解決法を示す．

Gathering Data in Dynamic Change of Sensor Network Topology

Yutaka Yanagisawa　Yasushi Sakurai　Yasue Kishino
Takuya Maekawa　Koji Kamei　Takeshi Okadome

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-Soraku, Kyoto, 619–0237

Abstract

We propose a technique to gather spatial correlated data for a wireless sensor network even if the topology of the network changes dynamically. Our technique can gather data from the wireless sensor network with the minimum communication cost. We adapt the clustering technique to gather spatial correlated data efficiently. Moreover, we show a solution to the problem in applying the gathering technique into a practical sensor network such that the coding system uses much memory to assign a code to a pair of the integrated data.

1　はじめに

データを収集するときの通信コストを最小化することは，センサネットワークの研究課題の中で最も重要な一つである．通信コストを最小化することは，センサノードの消費電力を最小化することにつながる [7]．その結果，センサネットワークのライフタイムを最大化し，メンテナンスコストの削減が可能になる．通信コストの削減の研究は，物理層，ネットワーク層，アプリケーション層の各層において，それぞれ行われている [1]．本稿ではこれらのうち，アプリケーション層において相関性のある複数のデータを集約することで，通信コストを最小化する方法について述べる．特に本稿では，雨量や温度，湿度といった，センサ間の空間的な距離と観測されるデータの類似度との間に相関性が見られるデータの集約方式 [10] について議論する．

対象とするセンサネットワークは，n 台のセンサノード (source node) と 1 台のシンクノード (sink node) を持つ．センサノードは，センサを用いて雨量や温度などの観測値を取得する機能と，データの計算処理を行う機能，および近隣のセンサノードと通信する機能を持つ．データの収集とは，ある時刻にそれぞれのセンサノードが取得した観測値を，無線通信機能を用いてすべてシンクノードまで集めることを言う．また「通信コスト」とは，ノード間でデータを送信するために必要な最小のビット数とノード間の距離の積で与える．データは送信

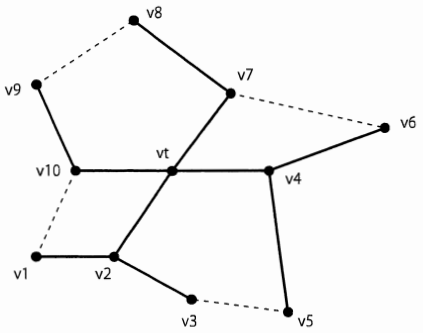

図 1: 対象とするセンサネットワークの構成図

時にエントロピー符号等を用いることでことで，冗長性を排除する．このとき，通信コストを最小化するとはすなわち，データ収集時のすべてのノード間の通信コスト（ビット数）の総和を最小化することである．以上の想定は，従来の通信コストを最小化する研究で対象としている，一般的なセンサネットワークの想定と同等である [4][10][8].

対象とするセンサネットワークの想定図を図 1 に示す．図の $v_1, \ldots, v_n$ はセンサノードである．v_t はシンクノードを示す．各センサノードは無線による通信が可能な距離に上限があるため，通信可能な隣接センサノードは限られる．多くのセンサノードはシンクノードとの直接の通信が行えない．データ収集のためには，他のセンサノードを経由してシンクノードまでデータを送る必要がある．

ここで，あるノード v_i が無記憶でランダムな情報源 X_i の生成するデータを観測するものとし，観測されたデータ送信するために必要な最小のビット数を $H(X_i)$ と記述する．また，隣接する二つのセンサノード v_i, v_j 間の送信路を $e(i,j)$ と記述し，$e(i,j)$ の通信コストを $w(i,j)$ と記述するものとする．v_i と v_t 間の通信路は $e(i,t)$，通信コストは $w(i,t)$ と記述する．このとき，図の v_4 のデータ収集のための通信コスト $w(4,t)$ は $H(X_4) \times len(e(4,t))$ と等しい (len は通信路の距離を与える関数である)．一方，隣接しない二つのノード v_2 と v_t の通信コスト $r(2,t)$ は $H(X_3) \times len(e(3,2)) + H(X_2) \times len(e(2,t))$ である．図 1 のネットワークでは，ノード v_3 のデータはノード v_2 を経由してシンクに送られるからである．なお $r(i,i) = 0$ であり，通信路が存在しないノードの組 v_k, v_l については $r(k,l) = 0$ とする．

以上の定義に基づけば，本稿で対象とする通信コストの最小化の課題は，次の式の値 W を最小化することであると定義できる．

$$W = \sum_{i=1}^{n} \sum_{j=1}^{n} r(i,j) \tag{1}$$

この課題は，二つの問題に分割できる．

1. W を最小化する各ノードからシンクまでの通信路の組み合わせを与える．
2. W を最小化するデータの符号化方式を与える．

この課題に対し，1) ネットワークトポロジーが変化しない，2) 予めセンサが観測する各値の生起確率分布が既知である，3) 観測データに空間的な相関性があるという三つの条件を満たすセンサネットワークにおいては，Cristescu [4] ら，および Liu [8] らにより通信コストを最小化する方法が示されている．これらの文献によれば，符号化方式として Slepian-Wolf Coding [12] を，通信路としては Shortest Path Tree [2] を用いることで，理論的に通信コストを最小化できることが示されている．じかし，実際のセンサネットワークにおいては，1), 2) の条件を満たすことは一般に困難である．

トポロジーに関しては，センサネットワークを構成する各ノードの脆弱性が要因となり，トポロジーの変化を生じさせることがある．ノードの通信能力は必ずしも信頼性が置けるものではない．バッテリや通信環境によって，一時的に通信が不可能となることがある．このとき，ノードが一時的にネットワークから欠損し，結果的にトポロジーの変化を生じさせる．メンテナンスによりそのノードが復帰した場合は，逆にネットワークにノードが追加されたこととなり，同様にトポロジーの変化を生じさせる．Silberstein [11] らによれば，森林の天候状況を観測するモニタシステムにおいて，ノードの故障や通信の失敗によりトポロジーの変化を余儀なくされるケースは 3% 程度であったという．筆者らの実験環境においても，100 個のセンサノードを稼働させた時の，ノードの欠損確率はおよそ 3% であった[1]．欠損が末端のノードで起こる場合は影響が少ないが，よりシンクノードに近いノードで発生すると影響が大きくなる．つまり，このようなトポロジーの動的な変化が起こった場合でも，通信コストを最小に近い値に維持できることが重要である．

次に，センサデータが観測する値の生起確率が既知であるという想定は，ほとんどのアプリケーションにおいて現実的ではない．たとえば，前述の森林の気温や湿度などを取得するためのセンサネットワークや，海洋の水

[1] ノードがデータを含むパケットの送信に失敗する確率はおよそ 30% であった．2 回のパケット再送を試みることを許すことで，実質的な欠損率は 3% 程度となる．

温や水流などを観測するセンサネットワークでは [6]，観測される値を得ることがそもそもの目的である．このようなアプリケーションで，あらかじめ値の生起確率が求められているのであれば，そもそもセンサネットワークを用いて観測する必要性がない．

さらに Slepian-Wolf Coding を用いると，各ノードが観測したデータに符号を割り当てる際に必要となる記憶領域のサイズがきわめて大きくなるという問題がある．具体的には，ノード数を n，観測されるデータの種類が m であるとき，記憶領域のサイズは最大で m^n となる．このため，例えば気温や雨量のように，連続した値を取るデータを観測するときには，必要となる記憶領域のサイズが極めて大きくなる可能性がある．

本稿では，以上の問題を解決することができる，センサネットワークのためのデータの集約方式について述べる．提案方式は，以下の三つの方式で構成される．

- **クラスタリング**: ノードを動的にクラスタ化し，クラスタごとにデータの集約を行う．そして集約したデータの組ごとに符号を割り当てる．これにより，トポロジーの動的な変化に対応する．
- **符号**: 過去に集約したデータの生起確率から，各ノードごとに独立して符号テーブルを作る．符号テーブルをノード間で交換する必要はない．これにより，生起確率が既知でなければならないという問題を回避する．
- **キャッシュ**: 記憶領域のサイズに上限を設定し，頻度推定手法によって得られた生起確率の高いデータのみをキャッシュに入れる．これらのデータのみにエントロピー符号を割り当て．生起確率が低いデータは静的な（通信コストの低減効果が得られない）符号を割り当てる．これにより，記憶領域のサイズが大きくなる問題を解決する．

このうち，クラスタリングと符号に関しては，従来手法を利用する．本稿ではキャッシュを導入することで，従来法による通信コストを削減する効果を大きく損なうことなく，記憶領域のサイズを小さく方法について主に述べる．これにより，効率的にデータ集約してデータ収集時の通信コストを低減できる．

2 課題

本章では，データ収集の対象となるセンサネットワークの構成と，対象とする課題の定義を行う．次に，このセンサネットワークにおける理論的な最小の通信コストについて述べる．

本稿で対象とするセンサネットワークの構成，および通信コストの最小化モデルについては，Cristescu ら [4][5] および Liu ら [8] が示したものと同等である．ここでは Liu らの文献の記述をもとに，対象とするモデルについて説明する．

図 1 に示すように，センサネットワークは n 個のセンサノード $v_1, \ldots, v_n$ とひとつのシンクノード v_t を含む．これらのノードの集合を $V = \{v_1, \ldots, v_n, v_t\}$ と記述する．通信が可能なセンサノードの対 v_i, v_j を結ぶ枝（通信路）を $e(i,j) \in E$ と表す．このとき，センサネットワークは，$n+1$ ノードをもつグラフ $G = (V, E)$ とみなせる．V の各ノードはデータを符号化する機能と，隣接するノードへデータを送信する機能をもつ．

あるノード $v_i \in V$ で観測するデータは，ランダムかつ離散的な情報源 X_i が発生させるものとする．離散的とは，観測される値の範囲が連続でない（たとえば $1, 2, \ldots, 10$ のように連続でない離散的な値）という意味である．$X_i, \ldots, X_n$ の結合ベクトル $X = \{X_i, \ldots, X_n\}$ は，同時確率分布 (joint probability distribution) $p(X_i = x_1, \ldots, X_n = x_n) = p(x_1, \ldots, x_n)$ で特徴付けられる．また τ を時刻とするとき，$X(\tau)$ は時刻 τ にすべてのノードにおいて同時に観測された観測値の結合ベクトルである．このとき，確率過程 $\{X(\tau)\}_{\tau=1}^{\infty}$ は定常過程である．$X_1, \ldots, X_n$ はそれぞれ独立同分布 (i.i.d.) であるとする．

通信路 $e(i,j)$ を流れるデータの 1 秒あたりの平均ビット数を $b(i,j)$ とし，$e(i,j)$ の通信コストを $r(i,j)$ とする．無線通信の場合，一般に $r(i,j)$ は記述の簡略化のため，以降 $r(i,j)$ を単に r と表記する．b, e も同様である．このとき，一般に通信コストは，通信路 e の距離と平均ビット数 b の積 $b \times len(e)$ で与えられる [5]．ここで len は通信路の距離を与える関数である．ただし，v_t との間に直接の通信路を持たないノードは，他のノードを経由してデータを送信しなければならない．各 $b(i,j)$ は，この中継されるデータすべてを含めた，$e(i,j)$ 上で通信するときに必要な平均ビット数である．なお，グラフ G は時刻に対して静的であり，いずれの時刻においてもノードの位置や通信路をもつノードの組み合わせは変化しないものとする．グラフのトポロジーが動的に変化する場合については後述する．

ここで，W_G をグラフ $G(V,E)$ の各通信路 $e \in E$ に関する通信コスト r の集合であるとするとき，このグラフの通信コストの総計 W_G は次式で表わされる．

$$W_G = \sum_{r \in R} r \qquad (2)$$

この W は式 1 と等価である．グラフ G が静的であるときは，この通信コストを最小化するデータ集約方式は，Liu らによって示されている．グラフ G の v_t を根とする Shortest Path Tree の各枝を通信路として用い，Slepian-Wolf Coding を用いて符号化することで，理論的に最小の通信コストでデータ収集を行うことができる．

ここで，グラフ G のノード v_t を根とする Shortest Path Tree をなす部分グラフを $G_{SPT}(V, E_{SPT})$ とする．このグラフ $G_{SPT}(V, E_{SPT})$ 上の二つのノード v_i, $v_j \in V$ について，これらの間を直接接続する枝 $e(i,j)$ がない場合，$e(i,j)$ という記述は $G_{SPT}(V, E_{SPT})$ 上の v_i, v_j の最短の経路（通信路）を表わす．また，ノード $v_1, \ldots, v_n$ は v_t からの距離が近い順にソートされているものとする．すなわち $len(e(1,t)) \leq len(e(2,t)) \leq \ldots \leq len(e(n,t))$ である．

このとき，グラフ G における通信コストの総和 W_G の下限は次式で与えられる．

$$W_G \geq \sum_{i=1}^{n} len(e(i,t)) \times H(X_i | X_{i-1}, \ldots, X_1) \qquad (3)$$

なお，グラフ G_{SPT} の通信コストの総和 $W_{G_{SPT}}$ の下限も式 3 と等しい．これらの証明は [4] [5] を参照されたい．$H(X_i)$ は定常情報源 X_i の平均情報量を表わす．

以上の条件のもとで，$X_1, \ldots, X_n$ が互いに完全に独立ではないとき，すなわちある時刻に各情報源が発生させる情報が互いに何らかの相関性を持つ場合に W_G を最小にできる（3 の右辺と左辺が等しくなる）データ集約法のひとつに Networked Slepian-Wolf Coding (NSWC) と呼ばれる方法がある [5]．NSWC が W_G を最小化できることに関する証明は [8] を参照されたい．

3 提案手法

提案するデータ集約手法は，1 章で述べたように三つの技術によって構成される．すなわち，ノードのクラスタリング法，符号テーブルの作成法，符号作成時に必要となる記憶領域のサイズの制限 (キャッシュ法) の三つである．本省では，これらの各技術についてそれぞれ 3.1, 3.2, および 3.3 節にて説明する．

3.1 クラスタリング

NSWC を行うノード数を減らすことは，NSWC を用いてデータに符号を割り当てる際に必要となる記憶領域のサイズを削減できる効果的な方法である．クラスタリングによりノード数を減らす方法については，Cristescu らによっても言及されている [4]．この方法は，グラフを複数のクラスタに分割し，クラスタ単位で NSWC を行う方式である．本稿では，最もシンプルなクラスタ化の方法を導入する．

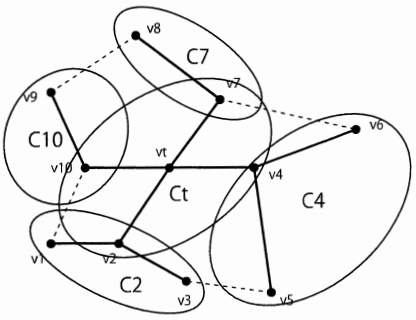

図 2: クラスタリングの例

クラスタに含まれるノードの数を少なくすることにより，符号の構築時に必要となる記憶領域のサイズそのものを小さくする効果と同時に，トポロジーが変化した場合に符号テーブルを書き換えるコストを大幅に削減する効果も期待できる．NSWC ではあるノード $v_i \in V$ がグラフ $G(V,E)$ から削除された場合，v_i より遠いノード $v_{i+1}, \ldots, v_n$ のすべてのノード間で使用する符号を作成しなおす必要．ノード v_i が v_t に近ければ近いほど，そのコストは大きくなる．

一方でクラスタ化することにより，全体の通信コストはネットワーク全体に対して NSWC を用いてデータを収集する場合よりも低下する．これは，異なるクラスタに含まれるノードの情報源の間に相関性があった場合，この相関性が通信量の削減に影響を与えなくなるためである．しかし，空間的な相関性をもつデータを収集する場合には，空間的に近接しているノードを同じクラスタに入れることで，高い相関性のある情報源をもつノードを同じクラスタに入れることができる．

ここで，グラフ G_{SPT} は定義から木状であるので，その部分木をクラスタとする方法が最もクラスタリングのためのコストが少なくなる．そのため，提案手法ではこの方法でクラスタリングを行う．

ノードの削除や追加の影響は，クラスタ内の木の深さが深くなるほど大きくなる．あるノードとそのノードに隣接し，なおかつそのノードよりシンクから遠いノード群をひとつのクラスタ C とする．クラスタはあきらかに

$G_{SPT}(V, E_{SPT})$ の部分木である．クラスタのうち一番シンクまでの距離が近いノードをクラスタヘッド (head) と呼ぶ．図 2 にクラスタ構成の例を示す．ノード $v_i \in V$ がクラスタヘッドであるようなクラスタを C_i と記述する．クラスタ C_i に含まれる，クラスタヘッド v_i 以外のノードを C_i の子ノードと呼ぶ．のクラスタ C_i は，明らかにグラフ G_{SPT} の連結な部分グラフである．すべての通信路が，いずれかのクラスタに含まれるようにクラスタリングを行う．このとき，$v_t \in G_{SPT}$ 以外でクラスタヘッドとなっている各ノード $v_i \in C_i$ は，自身がクラスタヘッドとなっているクラスタ C_i に属すると同時に，他のあるクラスタ $C_j : C_j \neq C_i$ にも含まれて．

以下，ノードの削除や変化に応じてクラスタを動的に作り直す方法について説明する．クラスタヘッドではないノードがグラフ G から削除されたときは，そのノードが含まれるクラスタが含むノードのみが符号を作成しなおすだけでよい．クラスタヘッドであるノードがグラフ G から削除された場合は，Shortest Path Tree を作成しなおす必要がある．このとき，最悪のケースで，削除されたノードがクラスタヘッドであったクラスタ C に含まれていたノードの数の 2 倍のクラスタが，クラスタ内で使う符号を作りなおす必要がある．これは，Shortest Path Tree を作成しなおしたときに C に含まれている各ノードがそれぞれ別のクラスタに所属することになる可能性があるためである．ただし，実際のセンサネットワークにおいては各クラスタに含まれるノード数は高々数個になることから，符号を作成しなおすのは高々数個のクラスタに限られる．

ノードの追加が行われた時も，削除時と同様に Shortest Path Tree を計算しなおし，それによってクラスタに所属するノードが変化したクラスタについてのみ符号を作成しなおす．

3.2 符号

提案手法では，NSWC をクラスタ毎に行う方法と，エントロピー符号と符号を一定時間間隔で更新する方式を組み合わせた方法の，二つの方法を用いる．ただし，NSWC は前述のように観測される値の生起確率が既知でなければならないため，実際のセンサネットワークへの適用が困難である．このため，本稿では性能の評価基準として利用するのみに留める．

NSWC を用いる方法　NSWC を用いる場合は，グラフ G_{SPT} 内のクラスタごとに独立した符号を用いる．すなわち，クラスタヘッドをシンクノードとみなしてシンクノードが復号を行い，子ノードは符号化したデータをクラスタヘッドに送信する．クラスタヘッド v_i が $v_i \neq v_t$ である場合は，二種類の符号を使用することになる．すなわち，v_i がクラスタヘッドとなっているクラスタ C_i 内で用いる符号と，v_i が子ノードとなっているクラスタの符号である．この場合，v_i は子ノードから受け取った符号を一旦復号し，これに v_i 自身が観測したデータを加えて符号化する．

このようにクラスタリングを行った場合，通信コスト W_G は下限値は式 3 と等しくなり，上限値は次式

$$\sum_{i=1}^{n} len(e(i,t)) \times H(X_i) \tag{4}$$

と等しくなる．情報源 $X_1, \ldots, X_n$ がそれぞれ完全に独立しているときは下限値となり，逆に互いに完全に従属しているときは上限値になる．このクラスタリング法は Cristescu らが示した一般的なクラスタリング手法を特殊化したものである．通信コストの詳しい議論については文献 [4] を参照されたい．

NSWC を用いない方法　NSWC を用いない場合は，各クラスタのひとつの子ノードとクラスタヘッドの対ごとに，ハフマン符号などのエントロピー符号を用いてデータの通信を行う．この方法を用いると，各クラスタごとの通信コストは式 4 と等しくなり，集約による通信コストの削減効果が全く得られない．しかし，グラフ全体で見た場合には通信コストの削減効果が現れる．グラフ G_{SPT} のクラスタヘッドの集合を $v_{H1}, \ldots, v_{Hm} \in VH$ とし，グラフ G_{HSPT} を G_{SPT} の部分木であり，なおかつ VH のみを含む最小のグラフとする．このクラスタヘッドのにみからなるグラフ G_{HSPT} に関しての通信コストの上限値と下限値は，グラフ G_{HSPT} をクラスタリングして NSWC を用いてデータ収集を行う場合と等しくなる．つまり，最も末端の（クラスタヘッドにならない）ノードが含まれるクラスタ内では通信コストを削減できないが，それよりシンクノードに近いクラスタ内では通信コストの削減ができる．

一方で，あるクラスタで用いる符号としてエントロピー符号をそのまま用いると，あらかじめ各観測値の生起確率が既知である必要が生じる．すると，NSWC と同様に実際のネットワークに適用することが難しいという問題が起こる．これを回避するため，過去に観測したデータ列から，各データの生起確率を推定する方法を用いる．ただし，2 で述べたように各情報源 $X_1, \ldots, X_n$ は時間的な相関性を持たないものとしている．このため，推定を時系列予測モデル等は用いて行うことはできない．

過去に観測したデータの頻度（ヒストグラム）を推定して，符号をクラスタ単位で一定の時間間隔ごとに作り直すという，最もシンプルな方法を用いる．

この方法で頻度推定を行って符号を構築しなおす場合，ネットワークトポロジーが変化しない場合においても，データ収集を行い始めた直後は通信コストが式 4 以上となる可能性がある．しかし，時間経過に従って式 4 以下，および式 3 以上の範囲内に収束する．収束の速度は $X_1,\ldots,X_n$ の各分布に依存する．トポロジーが変化する場合，変化が生じた直後には通信コストが式 3 を超える可能性がある．この場合でも，その後にトポロジー変化が生じない場合は収束するが，変化が生じ続ける場合は収束しない．通信コストの低減効率がどの程度悪化するかはトポロジー変化の頻度に依存する．変化の頻度が高いほど，通信コストの低減効果が得られなくなる．

以下，クラスタごとに異なる符号を用いる方法について述べる．あるクラスタヘッド v_{Hi} のクラスタ C_{Hi} について，j 個の子ノードをそぞれ $v_{i1},\ldots,v_{ij} \in VC_{Hi}$ と記述する．また v_{Hi} が子ノードとなっているクラスタが C_{Hk} であるとし，またこのクラスタ C_{Hk} のクラスタヘッドは v_{Hk} である．このとき，ノード v_{Hi} は各子ノードからデータを受け取る．集めたデータはそれぞれ $X_{i1},\ldots,X_{ij}$ である．クラスタ化しない NSWC ではこれらをそのまま v_{Hk} へ転送するが，本手法ではこれらを $X_{C_{Hi}} = \{X_{i1},\ldots,X_{ij}\}$ という形に合成してから通信路 $e(Hi,Hk)$ を経由して v_{Hk} へ転送する．

このとき，v_{Hi} が $X_{i1},\ldots,X_{ij}$ をそのまま転送する場合の通信路 $e(Hi,Hk)$ の通信コストの下限は $\sum_{l=1}^{j} len(e) \times H(X_{il})$ となる．一方で，合成したデータ $X_{C_{Hi}}$ を送る場合は $len(e) \times H(X_{i1},\ldots,X_{ij})$ となる．ここで，$X_{i1},\ldots,X_{ij}$ 間に相関性がある場合は $\sum_{l=1}^{j} H(X_{il}) \geq H(X_{i1},\ldots,X_{ij})$ である [12]．つまり，相関性が高いほ通信コストの削減効果が高くなる．

3.3 キャッシュ

ここまでに述べたクラスタ毎に異なる符号を用いる方式で，あるクラスタ C_{Hi} に関して X_{il} の発生させるデータの頻度推定を行うことを考える．個々の情報源 X_{il} がそれぞれ $s(X_{il})$ 種類の異なる値（シンボル）を発生させる場合，データをすべてカウントして頻度を得るためには最低 $\sum_{l=1}^{j} s(X_{il})$ の記憶領域が必要になる．さらに，それぞれの情報源が発生させるデータ間の相関性をすべて調べて $X_{C_{Hi}}$ の符号を生成するには，最低で $s(X_{C_{Hi}}) = \Pi_{l=1}^{k} s(X_{il})$ の記憶領域が必要となる．

クラスタに含まれるノードの数と情報源が発生させるシンボルの数によっては，$s(X_{C_{Hi}})$ は非常に大きくなる．このことは，実際のセンサノード上での実装を困難にさせる可能性がある．以下ではこの問題について，Frequent Item Counting [9][3] を用いて頻度推定を行い，使用すべき記憶領域のサイズを抑制する方法を示す．

あるクラスタヘッド v_{Hi} が子ノード $v_{i1},\ldots,v_{ik}$ をもち，各子ノードは情報源 $X_{i1},\ldots,X_{ij}$ が発生させるデータを v_{Hi} へ送るとする．このとき，v_{Hi} はさらにシンクに近いノードへデータを送信するときに，前述のようにシンボルの組を統合して送信する．この統合したデータは，情報源 $X_{C_{Hi}} = \{X_{i1},\ldots,X_{ij}\}$ が発生していると見なせる．このとき，符号を割り当てるためには $X_{C_{Hi}}$ が発生させる各シンボルの頻度を知る必要がある．しかし，前述のようにすべてのシンボルの頻度を漏れなく知るためには，$s(X_{C_{Hi}})$ だけの記憶領域を必要とする．この頻度推定に用いる記憶領域のサイズに上限を設ける方法を以下に示す．

- 記憶領域（キャッシュ）のサイズの上限値 ϕ を決める．ネットワークの管理者，あるいは設計者が自由な値に決めてよい．
- サイズ ϕ の記憶領域のみを用いて，Frequent Item Counting により，一定時間内に情報源 $X_{C_{Hi}}$ が発生させるそれぞれのシンボルについて頻度推定を行う．この結果，一定時間後には記憶領域の中には頻度推定に成功したシンボル（すなわち発生頻度の高いシンボル）のみが残る．これにより，頻度推定のための記憶領域のサイズは ϕ 内に抑えられる．
- 記憶領域内に残っているシンボルについてのみ，エントロピー符号を用いて符号化する．符号は一定の時間間隔ごとに，頻度推定の結果を利用して構築しなおす．記憶領域内にないシンボルについては，$X_{C_{Hi}}$ が発生させるシンボルは用いず，もともとの $X_{i1},\ldots,X_{ij}$ の各情報源が発生させたシンボルに対し，予めネットワーク全体で統一的かつ静的に各シンボルに割り当てた符号を使ってそれぞれ符号化する．このとき，エントロピー符号と静的な符号が同じ符号を割り当てないように，予め符号間で調整を行っておく．

頻度の推定精度は ϕ と各情報源が発生させるシンボルの種類数に依存し，これらはトレードオフの関係にある．ただし推定精度の下限は，たとえば文献 [9] のアルゴリズムであれば，ϕ とシンボルの種類数から予め計算により求めることができる．

キャッシュサイズが 0 のときは，いかなるシンボルに対してもエントロピー符号が割り当てられなくなるため，通信コストは理論的な上限値を超える可能性がある．一方，ϕ を無限大にするとすべてのシンボルに対してエントロピー符号が割り当てられることになる．このときは，十分に長い時間が経過してキャッシュ内のシンボルが情報源 $X_{C_{Hi}}$ が発生させるシンボルの生起確率順に並んだ状態になることで，通信コストは理論的な下限値と等しくなる．ϕ の値として 0 以上の有限な値を用いる場合は，通信コストは下限値以上の値になる．この値は ϕ とシンボルの種類数から計算可能であり，それぞれの情報源 $X_{i1},\ldots,X_{ij}$ が発生させるシンボルの生起確率が従う確率分布に依存する．今後の研究の中で，通信コストを計算によって求める方法について検討する．

4 まとめ

本稿では，トポロジーが変化しても通信コストを低く抑えることができるデータ収集方式について提案した．また通信コストの上限値と下限値についても議論を行った．今後は，シミュレーションによる評価と，実際のセンサネットワークへ適用して評価を行うことを予定している．

参考文献

[1] Kemal Akkaya and Mohamed Younis. A survey on routing protocols for wireless sensor networks. *Ad Hoc Networks*, 3:325–349, 2005.

[2] Robert S. Cahn. *Wide Area Network Design*. Morgan Kaufmann, 1998.

[3] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *ICALP '02: Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, pages 693–703, London, UK, 2002. Springer-Verlag.

[4] Razvan Cristescu, Baltasar Beferull-lozano, and Martin Vetterli. On network correlated data gathering. In *INFOCOM '04: Proceedings of the 23rd Conference of the IEEE Communications Society*, pages 2571–2582, 2004.

[5] Razvan Cristescu, Baltasar Beferull-lozano, and Martin Vetterli. Networked slepian-wolf: theory, algorithms, and scaling laws. *IEEE Transactions on Information Theory*, 51(12):4057–4073, 2005.

[6] Naomi Ehrich Leonard, Derek Paley, Francois Lekien, Rodolphe Sepulchre, David Fratantoni, and Russ Davis. Collective motion, sensor networks, and ocean sampling. *Proceedings of the IEEE, special issue on the emerging technology of networked control systems*, (95):48–74, 2007.

[7] S. Lindsey, C. Raghavendra, and K.M. Sivalingam. Data gathering algorithms in sensor networks using energy metrics. *IEEE Transactions on Parallel and Distributed Systems*, 13(9):924–935, 2002.

[8] Junning Liu, Micah Adler, Don Towsley, and Chun Zhang. On optimal communication cost for gathering correlated data through wireless sensor networks. In *MobiCom '06: Proceedings of the 12th annual international conference on Mobile computing and networking*, pages 310–321, New York, NY, USA, 2006. ACM.

[9] Gurmeet Singh Manku and Rajeev Motwani. Approximate frequency counts over data streams. In *VLDB*, pages 346–357, 2002.

[10] Sundeep Pattem, Bhaskar Krishnamachari, and Ramesh Govindan. The impact of spatial correlation on routing with compression in wireless sensor networks. In *IPSN '04: Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 28–35, New York, NY, USA, 2004. ACM.

[11] Adam Silberstein, Gavino Puggioni, Alan Gelfand, Kamesh Munagala, and Jun Yang. Suppression and failures in sensor networks: a bayesian approach. In *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, pages 842–853. VLDB Endowment, 2007.

[12] D. Slepian and J.K. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on information Theory*, 19(4):471–480, 1973.