2C-5

環境情報データベース向け高性能センサデータ圧縮方式

加藤　守†　　谷垣　宏一†　　郡　光則†

三菱電機株式会社　情報技術総合研究所†

1. はじめに

大量の電力量センサ・室内温度センサ等から収集する環境データ向けのデータ圧縮方式として、DGR (Difference, Gamma, Run-length)符号化方式を提案する。本方式は、時系列上の変化が緩やかで連続的であるという環境データの特徴を利用した予測符号化方式であり、速度要件の厳しい商用環境情報データベースへの適用において、実用的な圧縮性能を得ることができる。オフィス環境で収集された環境データを用いた評価実験の結果より、本方式の有効性を示す。

2. 背景と課題

改正省エネルギー法の施行により、環境・省エネへの取り組みが重要な経営課題と位置付けられるようになってきた。これに対し、我々は、電力・温度など省エネに関係するセンサデータをリアルタイムに収集し一元管理する環境情報データベース向けの高速集計検索エンジン AQL (Analytical Query Language)[1]の開発に取り組んできた。AQL は元々データウェアハウス向けに開発したものであり、ログデータなどでは品名・取引先名などのデータ項目毎に限られた種類のデータが出現することが多いという特性に着目した RID(Run-length, Index and Difference)符号化方式[2]によるデータ圧縮を採用している。一方、センサデータは、連続的に変化する数値データであり、ログデータとは異なる特性を持つため、センサデータの効率的な圧縮が課題であった。

3. AQL向けデータ圧縮の要件

AQL 向けのセンサデータ圧縮方式は、以下の要件を満たす必要がある。

(1) ストリーム独立処理： センサ単位にストリームデータの圧縮・伸張が可能であり、他のセンサの値に依存しないこと。

(2) 高速処理： 圧縮・伸張処理の速度が、ストレージのデータ転送速度と同程度に高速な計算で実現可能であり、一連のデータ処理の速度ボトルネックとならないこと。

(3) 高圧縮率： 上記要件を満たしつつ可能な限り高い圧縮率が得られること。

A High-performance Sensor Data Compression for Consolidated Electricity Consumption Database
Mamoru Kato, Koichi Tanigaki, Mitsunori Kori
Information Technology R&D Center, Mitsubishi Electric Corporation

4. DGR符号化方式

前述の要件を満たす圧縮方式として、DGR (Difference, Gamma, Run-length)符号化方式を提案する。本符号化方式は、予測符号化＋ガンマ符号化＋ランレングス符号化により構成する。

4.1. 予測符号化

計算量の面で有利な、直前の値との差分を符号化する方式とする。センサデータの観測値時系列を$(x_0, \cdots, x_t, \cdots)$とするとき、時刻 t の観測値 x_t の予測値として x_{t-1} を用い、予測誤差 $r_t = x_t - x_{t-1}$ を後述の方法で符号化する。

4.2. ガンマ符号化

予測誤差を可変長符号により符号化する。計算量を減らすためにガンマ符号を用いるが、ガンマ符号は正整数のみに対して定義されているため、予測誤差 r_t を正整数に写像して対応する符号を得る。整数全体に拡張したガンマ符号による予測誤差 r_t の符号語 ExGamma(r_t)は、元のガンマ符号 Gamma(r_t)を用いて次式により求める。

$$\mathrm{ExGamma}(r_t) \equiv \begin{cases} \mathrm{Gamma}(2r_t) & (r_t > 0) \\ \mathrm{Gamma}(2|r_t|+1) & (r_t < 0) \\ 1 & (r_t = 0) \end{cases}$$

4.3. ランレングス符号化

環境データの中には、例えば室内湿度のようにほぼ一定値を保つよう制御されているデータがある。そうした変化の少ないデータに対してはランレングス符号化が有効である。

予測誤差 r が n 回連続するとき、次の３方式のいずれかにより符号化を行う。これらの方式を対象データに適用し、実際に最も良い圧縮率が得られる方式に圧縮ブロック毎に切り換える。

- 方式 G： ランレングスを使わず、符号語 ExGamma(r)を n 回出力する。
- 方式 G0R： r＝0 の場合はランレングスを使い、ExGamma(r)と Gamma(n)を出力する。r≠0 の場合は方式 G 同様に ExGamma(r)を n 回出力する。
- 方式 G01R： r∈{-1, 0, +1}の場合はランレングスを使い、ExGamma(r)と Gamma(n)を出力する。それ以外の場合は、方式 G 同様に、

ExGamma(r)を n 回出力する。

5. 評価

本圧縮方式の有効性を検証するため、AQL への実装を行い、実際の環境データを用いて圧縮率および速度性能の評価を行った。

5.1. 評価データ

評価データとしてオフィス環境で収集した電力量、室内温度・湿度のデータを使用した（表1）。これらのデータは計測間隔 1 分で計測したデータであるが、計測間隔の違いが圧縮率に与える影響を見るために、10 分、60 分間隔でリサンプリングしたデータを作成した。

表 1　評価データ

種類	計測間隔	データ量
電力量	1 分	13.8MB（36 計測点×100,562 レコード×4 バイト整数）
室内温度	1 分	31.6MB（77 計測点×107,650 レコード×4 バイト整数）
室内湿度	1 分	12.3MB（30 計測点×107,164 レコード×4 バイト整数）

5.2. 評価環境

評価には表 2 に示す PC サーバを用いた。

表 2　評価環境

OS	Windows Server 2008 SP2 Standard 32 bit
CPU	Intel Xeon E5345 2.33GHz (Quad core)× 1
Memory	6 GB
HDD	台数：2 台(RAID1)，　回転速度：15,000rpm, 接続 I/F：SAS

5.3. 圧縮率

各データに本圧縮方式(DGR)を適用した際の圧縮率を従来方式(RID) [2]と比較した結果を図 1 に示す。但し、圧縮率は圧縮サイズ／非圧縮サイズとし、圧縮サイズはデータを AQL にロードしたときのファイルサイズ、非圧縮サイズは表 1 に示すように 1 計測値を 4 バイト整数とし、計測点数×レコード数×4 バイトで計算した。

本 DGR 方式により、オリジナル（1 分間隔）の湿度・温度・電力量データを 1%～5%に圧縮することができた。サンプリング周期を広げると本方式が仮定する時系列上の冗長性が減少するため、圧縮率も劣化する傾向が見られるが、60 分間隔まで広げた場合でも 7%～18%の圧縮率が得られる。従来の RID 方式と比べると 44%～77%に圧縮できている。

5.4. 速度性能

集計・検索の高速処理に重要な伸張速度について、従来の方式との比較を実施した。データは電力量（60 分間隔）を AQL に 2500 回繰り返しロードした後、ヒット数 0 となる検索問い合わせの実行時間を計測した。圧縮前のデータ量は 2.5 億レコード、37GB である。

伸張速度の比較を表 3 に示す。CPU コア当りの伸張スループットは 396MB/秒で、従来方式の 64%となり、実用的に充分な性能が得られることが確認できた。大規模ストレージの処理ではしばしば、本実験のようにディスク I/O がボトルネックとなって、CPU 使用率が 100%にならない。このような条件下の実行時間は、ディスクからの読み出しデータサイズに比例する。このため、データ圧縮率の高い本方式では、CPU 負荷が従来方式と比べ増加しているにもかかわらず、実行時間を短縮することができる。

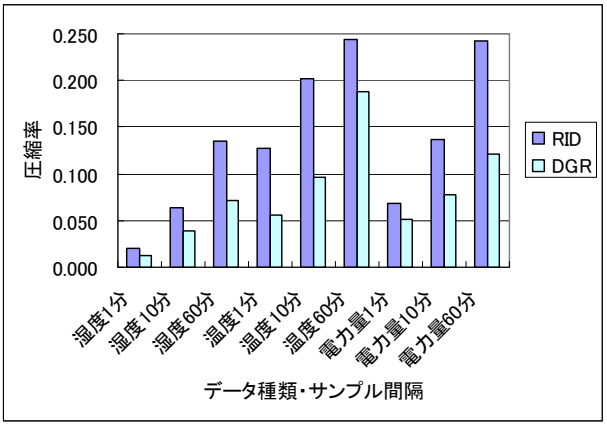

図 1　圧縮率比較

表 3　伸張速度比較

圧縮方式	圧縮率	実行時間	CPU 使用率	CPU コア当りの伸張スループット
DGR(提案)	11%	55 秒	44%	396MB/秒
RID	22%	111 秒	14%	617MB/秒

6. おわりに

環境情報データベース向けのセンサデータ圧縮方式を提案し、商用システム上に実装した。実データを用いた評価結果より本方式の有効性を確認した。

参考文献

[1] 山岸義徳,他：高速集計検索エンジンとセンサデータベースへの応用，三菱電機技報，Vol.83，No.12，pp.11-14(2009)
[2] 郡光則:データウェアハウス向け高性能データ圧縮方式，情報処理学会論文誌，Vol.47，No.SIG13，pp.58-73(2006)
[3] 竹田義聡,他：環境情報データベース向けリアルタイムセンサデータロード方式，情報処理学会第 73 回全国大会