D-033

データストリーム処理を高速・省メモリで行うためのスケッチ生成方式

A High-Speed and Small Memory Sized Sketching Method for Data Stream Processing

海老山 知生　喜田 弘司　藤山 健一郎　今井 照之　中村 暢達

Tomoo Ebiyama　Koji Kida　Ken-ichiro Fujiyama　Teruyuki Imai　Nobutatsu Nakamura

1．はじめに

IT 技術が社会に広く普及し，IC カードの読み取りデータや，GPS の位置データなどの小サイズのデータが大量・連続的に発生している．今後，これらのデータをリアルタイムに収集・分析することがますます重要になる．

データをリアルタイムに処理するためには，一旦蓄積してからバッチ的に処理する従来の計算モデルでは限界がある．そこで，我々は，データを蓄積せず，収集過程で流れ作業的に分散処理を行うデータストリーム処理方式の研究を進めている．データストリーム処理方式では，基本的には処理済のデータは破棄し，最新のデータのみをオンメモリで高速に処理する．しかし，過去のデータを完全に破棄してしまうと，過去のデータとの比較など，過去のデータが使った解析が一切できない．そのため，スケッチ[1]と呼ばれる過去のデータの要約を記憶しておき，必要に応じて利用する．本稿では，スケッチの生成方式について提案し，評価した結果について述べる．

2．スケッチ生成の要件

データ要約の性能指標はデータの要約率と要約の精度である．連続的に発生する時系列データを対象とする場合，データは発生し続けるため，すべてのデータが揃ってから要約する，ということができない．また，ある程度データが蓄積されてから要約しようとするとデータが蓄積されるまでのタイムラグによってリアルタイム性が損なわれてしまう．そのため，データが次々に発生する状況下で，データを蓄積することなく，データが発生するたびに逐次的に要約しなければならず，要約率や精度を上げることが困難である．このように，データが発生するたびに逐次的に要約しなければならない状況下で，高精度かつ少ないデータサイズに要約することがスケッチ生成の要件である．

3．スケッチ生成方式

3.1．スケッチ生成の概要

本研究で扱うデータは連続的に発生する時系列データであり，提案方式によるデータ要約は，発生したデータ列を複数の直線（１次関数）で近似することで要約し，データ量を削減する（図 1 参照）．

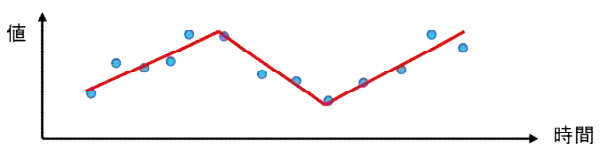

図 1　スケッチ生成の概要

NEC サービスプラットフォーム研究所
Service Platforms Research Laboratories, NEC Corporation

3.2．課題

提案方式では，要約の精度は，実際のデータの値と近似した直線との距離（誤差）が小さいほど高いと言える．また，要約率は，データを近似した直線の数が少ないほど高いと言える．すなわち，データを少ない数の直線で近似し，かつ，データの値と近似した直線との距離の和が小さくなるように要約率と精度のバランスを考慮して要約しなければならない．例えば，図 1 に示す例では 3 本の直線で近似しているが，1 本の直線で近似すれば，要約率は高くなる．しかし，精度が下がってしまう．逆に，より多くの直線で近似すれば精度は高くなるが，要約率が下がってしまう．

さらに，データは連続的に発生するため，データが発生するたびに，近似する直線を逐次的に決定する必要がある．

3.3．スケッチ生成のアルゴリズム

提案方式は，状態が確定した決定直線と，状態が確定しない暫定直線の 2 種類を用いてデータの要約を行う．暫定直線は，時間的に一番新しい直線であり，データが入力されるたびにそのデータを含めて最適な直線になるように修正される．新たなデータが暫定直線から大きく外れる場合は，新たな直線を生成し，暫定状態だった直線を決定直線にする．この操作をデータが入力されるたびに繰り返すことで逐次的に要約することが可能となる．

暫定直線は，新たにデータが入力されたとき，入力されたデータの実際の値（以降，実値）と近似直線から計算される値（以降，計算値）との差（図 2 参照）と予め設定しておくパラメータ T1，T2（詳細は後述）によって状態を変える．状態変化には以下の 3 パターンがある（図 3 参照）．

単純延長パターン：実値と計算値との差がパラメータ T1 より小さい場合は，暫定状態の近似直線を単純に延長する．

再計算パターン:実値と計算値との差がパラメータ T1 より大きく，パラメータ T2 より小さい場合は，暫定状態の近似直線と新たに入力されたデータとを使って直線の再計算（直線の補正）を行う．

新直線生成パターン：実値と計算値との差がパラメータ T2 より大きい場合は，新たな直線を引く．

○パラメータ T1 および T2

パラメータ T1 は，直線を単純に延長させるかどうかを決めるためのパラメータである．パラメータ T1 が大きくなるほど，実値と計算値との差が大きくても直線を単純に延長する．パラメータ T2 は，直線を再計算するかどうかを決めるためのパラメータである．直線の再計算を行うことによって，直線を単純に延長する場合に比べて要約の精度が高くなる．パラメータ T2 が大きくなるほど，実値と計算値との差が大きくても直線を再計算する．実値と計算値の差が T2 よりも大きくなった場合は，新たな直線を生成する．

以上のように，T1，T2 をより大きな値に設定するほど要約率が上がる代わりに精度が下がる．逆に T1，T2 を小さな値に設定するほど精度が上がる代わりに要約率が下がる．

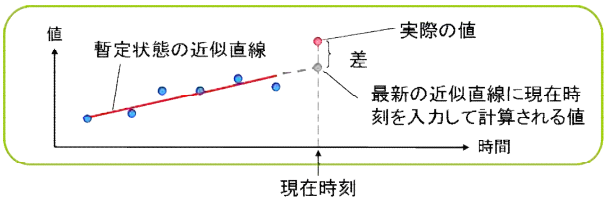

図 2　実際の値と計算値

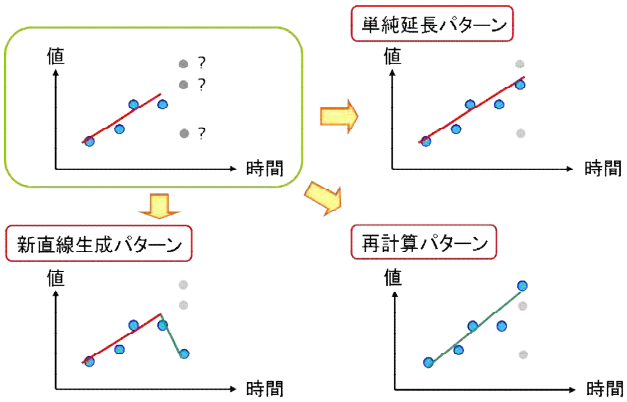

図 3　スケッチ生成のアルゴリズム

4．実験

4.1．実験の目的

本実験では，提案方式によって要約されたデータの要約率と要約の精度をパラメータ T1，T2 を様々に変化させながら評価することが目的である．

4.2．実験方法

本実験では，Web アクセスログから集計される Web ページのアクセス数を入力データとして用いた．

精度は，要約されたデータを用いて Web ページのアクセス数を集計し，正解データと比較することで計測した．

4.3．実験結果

データ要約率の評価：

パラメータ T1 および T2 の値を 0.5～10.0 まで 0.5 刻みで変えながら（ただし，T1 < T2 を満たす）計測した．要約後のデータサイズとパラメータ T1,T2 との関係をグラフ化したものを図 4 に示す．

今回の実験で使用した 24 時間分のログデータのうち，アクセス数の集計に必要な部分だけを集めたデータのサイズは約 100MB である．パラメータの値によってデータ要約率は異なるが，例えば，T1 = 2.0，T2 = 5.0 という値を設定した場合は，要約後のデータサイズは約 6.6MB であり，データをそのまま記憶する場合（100MB）に比べて，約 1/15 程度のデータ量に要約できている．

データ要約の精度の評価：

要約データは，パラメータ T1 および T2 の値を 0.5～10.0 まで 0.5 刻みで変えながら作成した．平均誤差とパラメータ T1,T2 との関係をグラフ化したものを図 5 に示す．

なお，誤差は以下の式により算出した．

「誤差(%)」＝ |(「要約データを用いて計算した値」－「正解の値」)|÷「正解の値」×100

パラメータの値によって誤差の大きさは異なるが，例えば，T1 = 2.0，T2 = 5.0 という値を設定した場合は，平均誤差が約 6.5%程度，最大誤差が約 20%程度である．

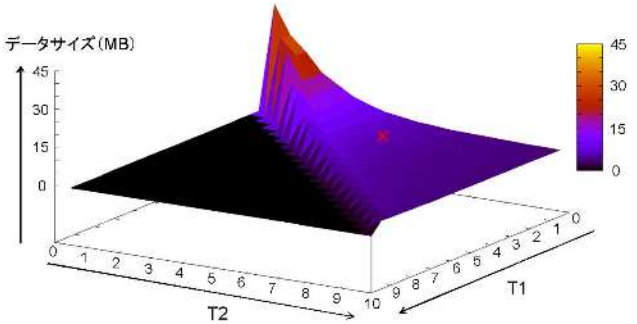

図 4　要約後のデータサイズ

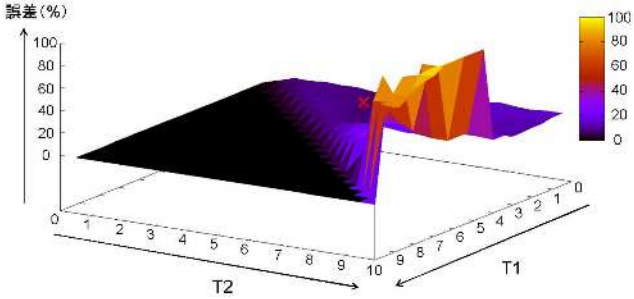

図 5　平均誤差

5．考察

要約率と精度はトレードオフの関係である．そのため，要約率と精度のバランスをとる必要がある．要約率と精度のバランスはパラメータ T1，T2 で調整することができ，今回の実験で使用したデータでは，T1 = 2.0，T2 = 5.0 と設定するとバランスが良くなる．このとき，要約後のデータ量は，データをそのまま記憶させたときのデータ量に比べて約 1/15 のサイズになる．また，要約データを用いたときの誤差は，平均で 6.5%程度，最大で 20%程度である．

大規模な Web サイトでは，1 日のログ発生量は数 G バイトにもなる．汎用 PC に搭載のメモリは 2G バイト程度であるので，データを要約しない場合，1 日分のログも分析できない可能性がある．しかし，本方式を用いることで約 2 週間分のログを分析することができ，より高度な統計処理が行える見込みを得た．また，要約に伴う誤差は平均で 6.5%程度であり，この程度の誤差であれば，アクセスのトレンド等を分析するには十分であると考えられる．

謝辞

本研究の一部は，総務省からの委託研究「ユビキタスサービスプラットフォーム技術の研究開発」の成果である．

参考文献

[1] Babcock, B., Babu, S., Datar, M. et al., “Models and Issues in Data Stream Systems”, in Proc. ACM PODS’02, pp.1-16, ACM (2002)