複数のウィンドウサイズに対応したリアルタイムバースト検出手法の提案

蝦名 亮平† 上原子 正利† 小柳 滋†
†立命館大学 情報理工学部

1 はじめに

バーストとはデータストリームにおけるイベントの異常な集中発生であり，オンラインニュース，ブログ，掲示板，通信記録などのデータストリームの特徴的な状態を解析するために重要である．これまで様々なバースト検出手法 [1, 2] が提案されているが，これらの手法は密度変化が大きいデータのリアルタイム解析には不向きである．本稿では密度の変化が大きいデータストリームのバーストをリアルタイムに判定する手法を提案する．

2 既存手法

本手法は Zhang らの手法 [2] の中で用いられる *Aggregation Pyramid* と呼ばれるデータ表現（図1）を参考にしている．これは N 個のレベルから成り，レベル h は

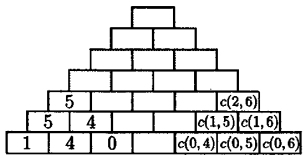

図1 Aggregation Pyramid の例

$N-h$ 個のセルを持つ．時間 t に終了するレベル h のセルを $c(h,t)$ と定義すると，$c(h,t)$ が持つ値は $c(0,t-h)$ から $c(0,t)$ が持つ値を利用して計算される．あるイベントの発生数について解析する場合，レベル0の各セルは観測された値（一定期間内のイベント発生数，期間の開始時間，期間の終了時間，期間の長さ）を持つ．$c(h,t)$ が持つ値は $c(0,t-h)$ から $c(0,t)$ が持つイベント発生数の合計，開始時間の最小値，終了時間の最大値，期間の長さの合計となる．また，$c(h,t)$ が持つ値は $c(h-1,t-1)$ と $c(0,t)$ を組み合わせた値になる．監視している期間の長さ，つまりウィンドウサイズが w のバースト判定の閾値を $f(w)$ とすると，ウィンドウサイズ w のセルの値が $f(w)$ 以上であればバーストが起きていると判定できる．

Zhang らの手法では一定期間内のイベントの発生数によってバーストを判定するため，長期間イベントが発生していなくても一定時間ごとに保持しているデータが更新されてしまい無駄な計算が行われる．また，その時点までに観測されたデータを基にして閾値が決定されることや，バースト検出が可能であるウィンドウサイズが一定であることから，データの密度が大きく変化すると適切に対応できない場合がある．本稿ではこれらの問題に対応したバースト検出手法を提案する．

3 提案手法

本手法は aggregation pyramid を利用するが，Zhang らの手法とは各セルに格納される値が異なる．各セルは合計到着間隔 *gaps*，到着時間 *arrt*，イベント到着数 *arrn* の3つのデータを持つ．

一連の $n+1$ 個のイベントが発生すると，各イベントの n 個の発生間隔を $x=(x_1,x_2,\ldots,x_n)$ と表す．また，解析したいウィンドウサイズの最小値 W_{min} を定義する．新しくイベントが発生すると，以下のルールに従って aggregation pyramid のレベル0に相当するセルにデータが格納される．

- $x_i \geq W_{min}$ の場合，$c(0,t).gaps$ には x_i を，$c(0,t).arrt$ には x_i が終了した時間を，$c(0,t).arrn$ には1を格納する．$i++$，$t++$ とする．
- $x_i < W_{min}$ の場合，$c(0,t).arrt$ には $c(0,t-1).arrt+W_{min}$ を，$c(0,t).arrn$ には $c(0,t-1).arrt$ より後から $c(0,t).arrt$ までの期間に発生したイベントの発生回数を，$c(0,t).gaps$ には W_{min} を格納する．$i=i+c(0,t).arrn$，$t++$ とする．i 更新後の x_i は $c(0,t).arrt$ から次のイベント発生までの経過時間に書き換える．

また，レベル1以上のセルは以下のようになる

- $c(h,t).gaps = c(h-1,t-1).gaps + c(0,t).gaps$
- $c(h,t).arrt = c(0,t).arrt$
- $c(h,t).arrn = c(h-1,t-1).arrn + c(0,t).arrn$

こうすることによって，データの更新がイベント発生時のみとなり，期間 W_{min} 内に複数のイベントが発生した場合はそれらの情報が1つのセルに圧縮される．

図2において，実線で書かれたセルは，色のついたセルが新しく生成されるときに保持していなければならないセルである．$c(h,t)$ の生成時に，対応するレベル0の

A Real-Time Burst Detection Method for Multiple Window Sizes
Ryohei EBINA†, Masatoshi KAMIHARAKO†, Shigeru OYANAGI†
†College of Information Science and Engineering, Ritsumeikan University

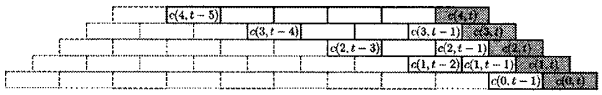

図 2　提案手法でデータの更新時に必要なセル

セルが重複していない最新の 2 つのセルである $c(h,t)$ と $c(h,t-h-1)$ を，後に示す方法で比較してバーストを判定し，結果に関わらず $c(h,t-h-1)$ を破棄する．ただし，$c(0,t-1)$ は $c(1,t)$ の生成に必要なため，$c(1,t)$ を生成後に破棄する．

本手法ではバーストが起きている期間を，到着間隔が重複していない直前の状態よりも急激に小さくなっている期間と定義する．各セルを同じ条件で比較するために 1 つのセル内の到着間隔の 1 つあたりの平均値を平均到着間隔 $c(h,t).gapa = c(h,t).gaps/c(h,t).arrn$ と定義し、$c(h,t).gapa$ と $c(h,t-h-1).gapa$ を比較する．バーストを判定するパラメータ $\beta(0 < \beta < 1)$ を用いて、$c(h,t).gapa \leq \beta \times c(h,t-h-1).gapa$ となるときバーストが起きていると定義する．また，バーストと判定できる最低イベント到着数 A_{min} を設定し，$c(h,t).arrn < A_{min}$ となるときは $c(h,t)$ のバースト判定を行わない．

4　評価実験

実データと人工データを用いて，既存手法と提案手法を比較する．提案手法のパラメータはそれぞれ $N = 50$, $\beta = 0.5$, $W_{min} = 1$, $A_{min} = 15$ とする．

実データとして新聞記事を用いる．図 3 は「CD-毎日新聞データ集 2006 年版」より，毎日新聞（2006 年発行分）の単語「サッカー」を含む記事数の変化と，提案手法によりバーストが検出された期間を表す．横軸は 2006 年 1 月 1 日からの経過日数，縦軸は単語「サッカー」が含まれた記事数を表す．この実験結果より，提案手法が

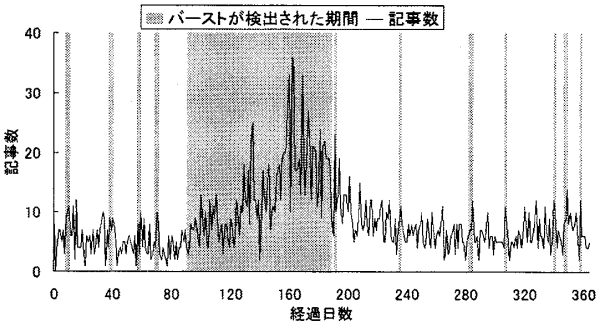

図 3　提案手法による実データの解析結果

実データの解析において有効であると言える．

次に，人工的に作成した密度変化の大きなデータを用いる．図 4 は Zhang らの手法によりバーストが検出された期間，図 5 は提案手法によりバーストが検出された期間を表す．横軸は時間，縦軸はイベント発生数を表す．既存手法の各パラメータは密度が高い期間

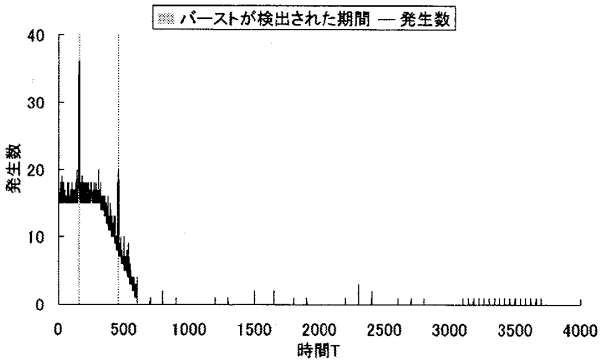

図 4　Zhang らの手法による人工データの解析結果

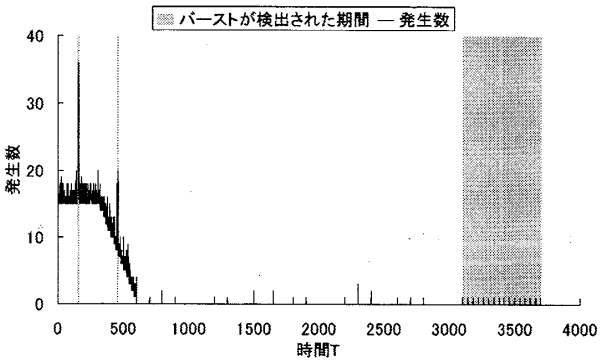

図 5　提案手法による人工データの解析結果

$(0 \leq T \leq 500)$ のバーストを正しく検出できるように設定した．しかし，密度が大きく下がった場合に適切にバーストを検出できなかった．一方，提案手法では密度が大きく変化してもパラメータを変更せずにバーストを検出することができた．

5　おわりに

本稿では，到着間隔が直前の状態よりも急激に狭くなっている期間を発見することにより，バーストを検出する手法を提案した．今後の課題として，使用メモリ量や計算量の削減が挙げられる．

参考文献

[1] Kleinberg, J.: Bursty and hierarchical structure in streams, *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, ACM, pp. 91–101 (2002).

[2] Zhang, X. and Shasha, D.: Better Burst Detection, *ICDE '06: Proceedings of the 22nd International Conference on Data Engineering*, Washington, DC, USA, IEEE Computer Society, p. 146 (2006).