4N-04

リアルタイムバースト検出手法における動的パラメータ決定法の提案

†福崎　達也　　‡中村　健二　　†小柳　滋
†立命館大学情報理工学部　　‡大阪経済大学情報社会学部

1　はじめに

データストリームの急な変化を素早く知る方法として，イベントの発生をリアルタイムに検出する手法が注目されている．データストリームにおけるイベントの集中発生はバーストと呼ばれ，蝦名らはバーストをリアルタイムに検出する手法を提案している [1] [2]．しかし，リアルタイムバーストの検出には複数のパラメータが存在し，観測対象ごとに正しいパラメータを設定する必要がある．観測対象ごとにバーストを観測できるパラメータは異なり，パラメータを動的に設定することで全てのデータを観測できることが望まれる．本研究ではリアルタイムバースト検出におけるイベントの発生間隔に着目して，パラメータを動的に設定する方法を提案する．

2　基本手法

リアルタイムバースト検出手法の概要を述べる．リアルタイムバーストはイベントの発生ごとに解析する．また，イベントの到着間隔が重複していない直前の状態よりも急激に短くなっている期間をバーストと定義する．より詳しいリアルタイムバーストの検出手法は [1] [2] を参照されたい．

2.1　データ構造

リアルタイムバースト検出手法のデータ構造を述べる．データ構造は *AggregationPyramid* と呼ばれるものを用いる．最新のデータが追加されると各段（レベル）の右側に追加される．ピラミッドは N 個のレベルから成り，時間 t に終了するレベル h のセルを $c(h,t)$ と定義する．レベル 0 は N 個のセルを持つ．レベル h は $N-h$ 個のセルを持つ．$c(h,t)$ 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．

各セルは合計到着間隔 *gaps*，到着時間 *arrt*，間隔個数 *gapn* の 3 つのデータを持つ．$n+1$ 個のイベント数に対して，各イベントの発生間隔を $x=(x_1,x_2,...,x_n)$ と表す．また，一度に大量のイベントが集中発生すると更新回数が膨大となるため，イベント数を一つのデータに圧縮する *Wmin* を設定する．本研究では $Wmin=1$ に固定する．

1. レベル 0 のセルの生成方法

   $x_i \geq Wmin$ の場合，$c(0,t).gaps = x_i$，
   $c(0,t).arrt = i+1$ 番目のイベントが発生した時刻
   $c(0,t).gapn = 1$, $i++$, $t++$,
   $x_i < Wmin$ の場合，
   $c(0,t).arrt = c(0,t-1).arrt + Wmin$,
   $c(0,t).gapn = c(0,t-1).arrt$ から $c(0,t).arrt$ 直前までの期間に発生したイベントの発生回数，
   $c(0,t).arrt$ にイベントが発生していなければ
   $c(0,t).gapn = c(0,t).gapn - 1$,
   $c(0,t).gaps = Wmin$, $i = i + c(0,t).gapn$, $t++$,
   $x_i = c(0,t).arrt$ から次のイベント発生までの経過時間，
   $c(0,t).arrt$ で複数イベントが発生していれば $x_i = 0$

2. レベル 1 以上のセルの生成方法

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

こうすることでデータの更新がイベント発生時のみとなる．

2.2　バースト判定方法

バーストの判定は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求める平均到着間隔関数を $avg(c(h,t)) = c(h,t).gaps/c(h,t).gapn$ のように定義し，$avg(c(h,t))$ と $avg(c(N-1,t-1-h))$ を比較する．$c(N-1,t-1-h)$ が存在しない場合は直前の期間で $c(h,t)$ と重複しない最高レベルのセルを比較する．今後，$c(h,t)$ に対する比較対象のセルは *tgcell* とする．パラメータ β を用いて，$avg(c(h,t))/avg(tgcell) \leqq \beta$ を満たすとき，バーストが発生していると定義する．

3　判定に用いるパラメータと動的パラメータ決定法

判定に用いるパラメータとそれらを動的に決定する方法を述べる．

3.1　パラメータ N

図 1 のようにバーストの解析には全てのレベルを用いるわけではなく，パラメータ N で指定した値のレベルのセルを比較対象とする．

このパラメータの動的な決定方法は，$N=50$ とする．N が 50 以上になると *tgcell* は古いデータを保持する

Dynamic Assignment of Parameters for Real Time Burst Detection
†Tatsuya FUKUZAKI　‡Kenji NAKAMURA　†Shigeru OYANAGI
†College of Information Science and Engineering, Ritsumeikan University
‡Faculty of Information Technology and Social Science Osaka University of Economics

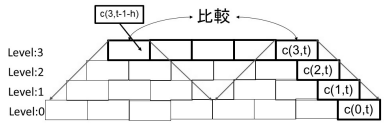

図 1: $N = 4$ で $c(3,t)$ が生成されたときに $c(N-1, t-1-3)$ と比較して，$c(3,t)$ がバーストしているかを判定する例

ことになるが，その必要はないためである．生成されるセル $c(h,t)$ が $t < 50$ であれば $N = t$ とする．

3.2 パラメータ *Amin*

$c(h,t)$ で突発的に少ない数のイベントが連続して発生した場合，過剰なバーストを避けるため $c(h,t).gapn < Amin$ となるときはバーストを判定しない．

このパラメータの動的な決定方法は，バーストしていない期間のイベント数の最大値とする．ただし，バースト観測を行わない最低値がある程度必要である．

3.3 パラメータ *Cmin*

バースト監視の初期段階において，比較対象のセル内の情報が極端に少ない場合，過剰にバーストが判定されてしまう．そこで $Cmin$ を用意し，$tgcell.gapn < Cmin$ のとき，$(avg(c(h,t))/avg(tgcell)) \times (Cmin/tgcell.gapn) \leqq \beta$ を満たせばバーストとする．

このパラメータの動的な決定方法は，$c(h,t)$ を判定ごとに $Cmin = tgcell.gaps$ とする．比較対象のセルのイベント発生間隔に対して合計到着間隔が大きいと，信頼度は低いと考えられるからである．ただし $Cmin < N$ とする．

3.4 パラメータ *Wmax*

データ構造上，長期間のバーストが観測されてしまう場合がある．そのため，$c(h,t)gaps > Wmax$ のとき $(avg(c(h,t))/avg(tgcell)) \times (c(h,t).gaps/Wmax) \leqq \beta$ を満たせばバーストとする．

このパラメータの動的な決定方法は，長期間のバースト観測を防ぐために $Wmax = 1$ とするが，イベント数が 1 日に必ず数件以上起こるものについては $Wmax$ を設定しないほうが良い．例えば，ニュース記事において「サッカー」や「経済」など定常的に使用されるキーワードで解析する場合である．

4 実験

動的なパラメータの設定によって，リアルタイムバーストが正しく観測できる一例を示す．実験に用いるデータは「CD-毎日新聞データ集 2006 年版」より，2006 年発行分の記事の本文キーワードと日付情報を用いる．動的に設定したパラメータでバーストを観測した結果を「torino2006」というキーワードでの 100 日間の記事数の推移で確認する．既存手法で設定されたパラメータ $N = 50$, $Cmin = 15$, $Amin = 15$, $Wmax = 1$ による観測結果を図 2 に示し，提案手法の動的設定パラメータによる観測結果を図 3 に示す．

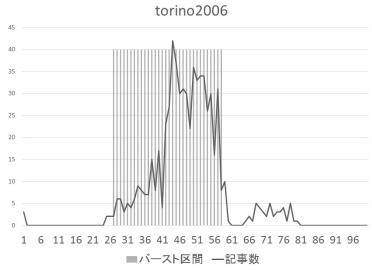

図 2: 既存手法による新聞記事データの解析結果

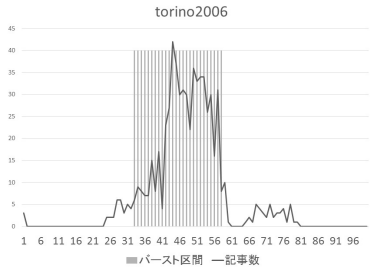

図 3: 提案手法による新聞記事データの解析結果

図 2 の既存手法のパラメータ設定での解析は記事数が少ない 28〜32 日目がバーストしているのに対し，図 3 の提案手法では 33 日目からバーストしていることがわかる．これはパラメータ $Cmin$ が動的に設定され，$Cmin$ が既存手法より有効に働いているからである．

5 おわりに

本研究でリアルタイムバースト検出におけるパラメータを動的に設定することができた．今後の課題として，動的なパラメータの設定によってバーストの誤検出が見られることもあるので，新聞記事以外の様々なデータを使って解析し，改善する必要がある．

参考文献

[1] 蛯名 亮平，中村 健二，小柳 滋．”リアルタイムバースト検出手法の提案”，日本データベース学会論文誌，Vol.9，No.2，2010 年

[2] 蛯名 亮平，中村 健二，小柳 滋．”リアルタイムバースト解析手法の提案”，情報処理学会論文誌，データベース，Vol.5，No.3，2012 年