情報処理学会論文誌　データベース　Vol.5　No.3　86–96 (Sep. 2012)

リアルタイムバースト解析手法の提案

蝦名 亮平[1,a)]　中村 健二[2]　小柳 滋[3]

受付日 2012年3月20日, 採録日 2012年5月9日

概要：データストリームにおけるイベントの異常な集中発生はバーストと呼ばれ，複数のウィンドウサイズにわたったバーストを解析する様々な手法が提案されている．本論文では，データストリームのバーストの度合いやランキングをリアルタイムに解析する手法を提案する．本手法では，イベントが発生していない期間の無駄なデータの更新を防ぎ，イベントが集中発生している期間のデータを圧縮して保持することにより，バースト検出の計算量を削減している．さらに，バーストの検出と同時にバーストの度合いやランキングを算出することにより，これらに基づいた効率的な解析が可能となる．実験により，解析結果の有効性を示す．

キーワード：バースト，データストリーム，リアルタイム，データマイニング，アルゴリズム

A Proposal for a Real-time Burst Analysis Method

Ryohei Ebina[1,a)]　Kenji Nakamura[2]　Shigeru Oyanagi[3]

Received: March 20, 2012, Accepted: May 9, 2012

Abstract: A burst is an unusually large number of events occurring within a certain period of time. Various burst analysis methods over multiple window size have been proposed. This article proposes a method to analyze a burst on real time. The proposed method reduces computation by avoiding redundant data updates of no event occurrence, and by suppressing data within a certain period in the case of emergent increase of events. In addition, a level and a rank of a burst can be analyzed simultaneously. The effectiveness of the proposed method is evaluated by experiments with real data.

Keywords: burst, data stream, real-time, data mining, algorithm

1. はじめに

ネットワーク技術の発展により大規模なデータストリームが登場し，これらの異常な状態を機械的に検出する方法が必要となっている．データストリームとは高速に次々と流れてくる大量のデータを指し，具体例としてオンラインニュース，ブログ，掲示板，通信記録，センサデータなどがあげられる．このデータストリームの異常な状態や急激な変化を素早く知る方法として，イベントの異常な集中発生をリアルタイムに検出する手法が期待されている．たとえば，オンラインニュースの急激な変化をリアルタイムに検出できれば，注目されている事柄などを膨大な情報の中から素早く把握可能となる．また，通信記録をリアルタイムに監視することによって，DoS 攻撃やアクセス状況の変化などの検出が可能となり，それらのアクセスに対し早急に対応可能となる．

データストリームにおけるイベントの異常な集中発生はバーストと呼ばれ，複数のウィンドウサイズにわたったバーストを解析する様々な手法 [1], [2], [3], [4], [5] が提案されている．また，バースト解析手法はブログ解析 [6], [7]，クラスタリング [8], [9]，検索 [10]，パーソナライゼーション [11] など様々な分野で応用されている．

[1] 立命館大学大学院情報理工学研究科
Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga 525–8577, Japan
[2] 大阪経済大学情報社会学部
Faculty of Information Technology and Social Sciences, Osaka University of Economics, Higashiyodogawa, Osaka 533–8533, Japan
[3] 立命館大学情報理工学部
College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga 525–8577, Japan
[a)] cs001066@ed.ritsumei.ac.jp

本論文では，データストリームのバーストをリアルタイムに解析する手法を提案する．これは，イベントが発生していない期間の無駄なデータの更新を防ぎ，イベントが集中発生している期間のデータを圧縮して保持することにより，リアルタイム処理を実現する．また，本手法ではバーストの詳細を示す指標としてバーストの強さを表すバーストの度合いと，異なるバーストを比較するための指標としてバーストの重さを解析して出力する．これにより，バーストの強さや期間の異なるバーストどうしを比較でき，効率的な情報抽出が可能となる．

本論文の構成は次のとおりである．2 章では関連研究について述べる．3 章ではリアルタイムにバーストを解析する手法を提案する．4 章では評価実験を行い，解析結果の有効性を示す．最後に 5 章で結論と今後の課題を述べる．

2. 関連研究

Kleinberg [1] は，テキストストリームのバーストをモデリングし，構造を抽出する方法について議論している．この手法は無限状態オートマトンを用いてストリームをモデリングすることをベースとしている．Kleinberg の手法の利点は，各トピックにおけるバーストの期間，度合い，重さを表すことができる点である．そのため，利用用途が広く，様々な応用研究で利用されている．しかし，あるイベントの発生に対して即座に解析することを前提としていないため，リアルタイムなバースト検出には不向きである．

Zhu と Shasha の手法 [2], [3], Zhang と Shasha の手法 [4] は複数のウィンドウサイズにわたったバーストを効率的に監視するバースト検出アルゴリズムを提案している．これらの手法は，監視する間隔を短くすることでリアルタイムに近いバースト検出が可能である．しかし，そのためには一定期間ごとにイベントの発生数を監視する必要があり，長期間イベントが発生していなくてもデータを更新するため，無駄な計算が行われる．このため，ドキュメントストリームに含まれるすべての名詞など，出現頻度にばらつきがある膨大な種類のイベントをリアルタイムに監視する場合，監視対象であるイベントの数に比例して計算時間が増加するため，効率的ではない．

リアルタイムにバーストを検出する手法として，文献 [5] が提案されている．これは，イベントが発生するごとにバーストしているかどうか判定することによって，リアルタイム検出を行う．また，イベントの集中発生時に，保持するデータを圧縮することで計算量を抑え，リアルタイム性の高いバースト検出を実現している．しかし，この手法はバーストがどれくらいの度合いで起こっているかや，バーストのランキングを決めるための評価指標がない．

このように，既存手法は大量のデータからリアルタイムにバーストを検出可能であるが，リアルタイムにバーストの度合いや，異なるバーストをランキングするための指標であるバーストの重さを算出する手法は提案されていない．本研究では文献 [5] と同様のデータ構造を用いてバーストをリアルタイムに検出し，さらに，そのバーストの度合いを解析し，重さによるランキング付けを行う．

3. 提案手法

本章では，ウィンドウサイズを固定せず，リアルタイムにバーストを解析する手法を提案する．

3.1 概要

提案手法では，一定時間ごとにバーストを解析するのではなく，イベント発生ごとにバーストを解析するため，イベントが発生していない期間の無駄な計算を防ぐことができる．また，一定期間内に複数のイベントが発生した場合は，それらを 1 つのデータに圧縮してバーストを解析することにより，計算回数の増加を防ぐ．本手法では，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して解析する．バーストの度合いは，到着間隔の変化率を用いて評価する．また，バーストの重さは，到着間隔の変化率がどれだけ正常範囲を超えているかを数値化することで算出する．

本章では提案手法の基礎となるデータ構造について述べ，その後，バーストの判定方法，度合いと重さを評価する方法について述べる．そして，バースト解析精度を向上させるための結果の補正方法について述べる．

3.2 データ構造

本手法のデータ構造は，Zhang と Shasha の手法 [4] で提案された aggregation pyramid を参考としている．Aggregation pyramid は Zhang と Shasha の手法のフレームワークの基となるデータ構造である．本節では最初に aggregation pyramid について述べ，次に提案手法のデータ構造について述べる．

3.2.1 Aggregation Pyramid

Aggregation pyramid は，**図 1** のようなデータ構造である．ここでは，最新のデータがセルに格納されて各レベルの右側に追加され，追加された分だけ各レベルの左側のセルが破棄される．**図 2** のように，aggregation pyramid は N 個のレベルからなり，時刻 t に終了するレベル h のセルを $c(h,t)$ と定義すると，以下のルールにより構成される．

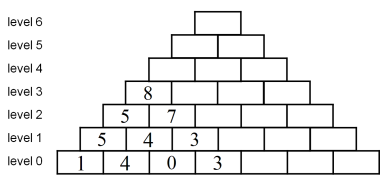

図 1　Aggregation pyramid の例
Fig. 1　Example of aggregation pyramid.

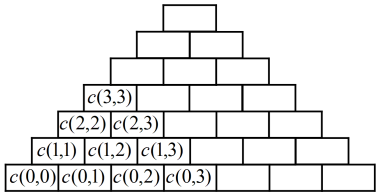

図 2　Aggregation pyramid のセルの表現方法
Fig. 2　Notation of cells in aggregation pyramid.

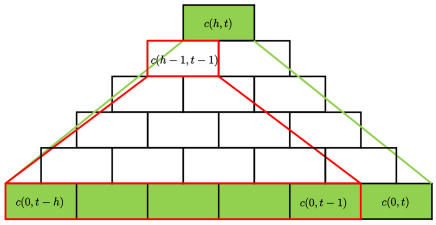

図 3　Aggregation pyramid の性質
Fig. 3　Feature of aggregation pyramid.

- レベル 0 は N 個のセルを持つ．これらは実際のデータに対応する．
- 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．
  - レベル 1 は $N-1$ 個のセルを持つ．レベル 1 の 1 番目のセルはレベル 0 の 1 番目と 2 番目のセルが持つデータを利用して計算する．レベル 1 の 2 番目のセルはレベル 0 の 2 番目と 3 番目のセルが持つデータを利用して計算する．これよりも高いレベルのセルも同様に計算する．
  - 最も高いレベル，つまりレベル $N-1$ は 1 つのセルを持つ．これはレベル 0 のすべてのセルの情報が集約されている．

$c(h,t)$ が持つ値の計算に必要なセルは $c(0,t-h)$ から $c(0,t)$ までのセルであるが，**図 3** で示すとおり，aggregation pyramid では，$c(0,t-h)$ から $c(0,t-1)$ までのセルの値は $c(h-1,t-1)$ に集約されるため，$c(h,t)$ が持つ値は $c(h-1,t-1)$ と $c(0,t)$ を利用して計算が可能である．

3.2.2　提案手法のデータ構造

本手法は aggregation pyramid の性質を応用した新たな手法である．各セルは合計到着間隔 $gaps$，到着時刻 $arrt$，間隔個数 $gapn$ の 3 つのデータを持つ．

一連の $n+1$ 個のイベントに対して，各イベントの n 個の発生間隔を $\mathbf{x}=(x_1,x_2,\ldots,x_n)$ と表す．また，大量のイベントが集中発生するとセルの更新回数が膨大となるため，解析するウィンドウサイズの最小値 W_{min} を定義する．W_{min} は，指定された値よりも短い期間内の情報を 1 つのセルに圧縮し，更新回数を一定以下に抑えるための値である．

新しくイベントが発生し新たな x_i を得ると，以下のルールに従って新たなセルを生成し，データ構造を構築する．

(1) レベル 0 のセルの生成方法
  (a) $x_i \geq W_{min}$ 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
    - $c(0,t).gapn = 1$
    - $i = i+1$
  (b) $x_i < W_{min}$ の場合
    - $c(0,t).arrt = c(0,t-1).arrt + W_{min}$
    - $c(0,t).gapn = c(0,t-1).arrt$ 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数
    - もし $c(0,t).arrt$ にイベントが発生していなければ，
      $c(0,t).gapn = c(0,t).gapn - 1$
    - $c(0,t).gaps = W_{min}$
    - $i = i + c(0,t).gapn$
    - $x_i = c(0,t).arrt$ から次のイベント発生までの経過時間．
      もし $c(0,t).arrt$ で複数の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
- $c(h,t).gaps = c(h-1,t-1).gaps + c(0,t).gaps$
- $c(h,t).arrt = c(0,t).arrt$
- $c(h,t).gapn = c(h-1,t-1).gapn + c(0,t).gapn$
- 最上位セル（$h = N-1$）の場合，$t = t+1$

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，イベントが発生していないときの無駄な計算を削減できる．また，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ため，膨大な量のイベントが集中発生したときにデータの更新回数を一定に制限できる．

本手法のデータ構造の構築例を以下に示す．$W_{min}=1$ と設定し，1 つのセルに格納されるデータを $c(h,t) = (gaps, arrt, gapn)$ と表現する．イベント発生時刻列 $\{0,1,6,6,6,6,6,9,9,10\}$ を得たとき，発生間隔列は $\{1,5,0,0,0,0,3,0,1\}$ となる．セルには到着間隔を基準にデータが格納されるが，**図 4** のように補正されてから圧縮されてセルに格納される．レベル 0 のセルは $c(0,0) = (1,1,1)$，$c(0,1) = (5,6,1)$，$c(0,2) = (1,7,4)$，$c(0,3) = (2,9,1)$，$c(0,4) = (1,10,2)$ の順に生成される．時刻 6 に 5 つのイベントが同時に発生している．期間 W_{min} の間にイベントが複数回発生しているため，時刻 6 から時刻 7 までの期間に 4 つの同じ長さの到着間隔が存在すると仮定して，その期間に対応する情報は圧縮されて 1 つのセルに格納される．これは，W_{min} よりも短い期間内での大量のイベントの発生などによって，生成されるセルの数が増加しすぎることを防ぐ．

しかし，こうすることによって実際に時刻 7 にイベント

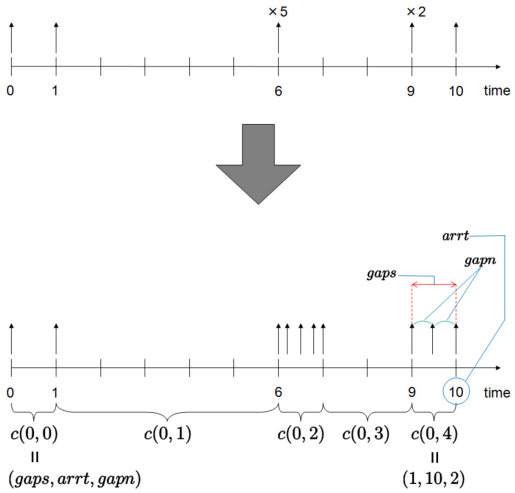

図 4　レベル 0 のセルの生成方法の例（$W_{min} = 1$）
Fig. 4　Example of generating cells at level 0 ($W_{min} = 1$).

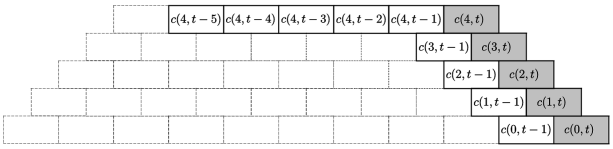

図 5　データ更新時に使用するセル（$N = 5$）
Fig. 5　Utilized cells for update process ($N = 5$).

が発生していなくても，$c(0,2).arrt$ は 7 に補正され，次のイベント発生時刻 9 までの到着間隔が実際には 3 であるが，2 として計算されてしまう．これは，解析対象とするウィンドウサイズの最小値 W_{min} を設定しているため，誤差が W_{min} 以下であれば問題ないと考える．

このように，到着間隔が長ければイベントが発生するたびにセルが生成される．到着間隔が短くなると許容誤差の範囲で情報が補正され，圧縮される．

図 5 において，実線で書かれたセルは，塗りつぶされ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．$c(h,t)$ の生成後に $c(h-1,t-1)$ を破棄する．対応するレベル 0 の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 $c(N-1,t-1-h)$ を 3.3 節に示す方法で比較してバーストを判定する．

3.3　バーストの解析方法

本手法では，新たに生成したセルと，到着間隔が重複していない直前の最上位レベルのセルを比較してバースト解析を行う．通常，$c(h,t)$ と比較されるセルは，$c(N-1,t-1-h)$ である．しかし，バーストを監視し始めた初期段階において，保持しているデータが少なく，新たに生成した $c(h,t)$ に対して $c(N-1,t-1-h)$ が存在しない場合は，期間が重複していないセルの中で最新かつ最もレベルが高いセルと比較してバースト解析を行う．たとえば，**図 6** において，新たに $c(0,2)$ が発生すると，$c(1,1)$ と比較し解析を行う．以後，$c(h,t)$ と比較されるセルを $tgcell$ と表現する．

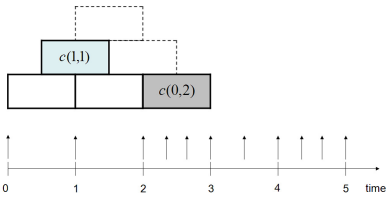

図 6　初期段階の比較対象のセルの例
Fig. 6　Example of compairing cells at an early stage.

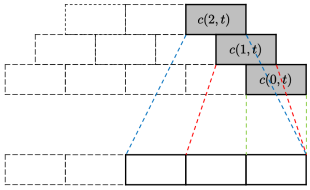

図 7　各期間のバースト度合いの決定方法
Fig. 7　Method of deciding a level of a burst.

3.3.1　バーストの判定基準

本手法では，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，バーストが発生し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算出する平均到着間隔関数 $avg(c(h,t))$ を式 (1) のように定義する．

$$avg(c(h,t)) = \frac{c(h,t).gaps}{c(h,t).gapn} \tag{1}$$

さらに，$avg(c(h,t))$ と $avg(tgcell)$ の平均到着間隔の変化率 $brt(c(h,t))$ を式 (2) のように表現する．

$$brt(c(h,t)) = \frac{avg(c(h,t))}{avg(tgcell)} \tag{2}$$

パラメータ β（$0 < \beta < 1$）を設定し，条件式 (3) を満たすとき，バーストが発生していると定義する．

$$brt(c(h,t)) \leq \beta \tag{3}$$

3.3.2　バーストの度合いの算出方法

バーストの度合いとは，バーストの強さを表す指標である．この指標を用いることで，バーストの強弱を判定できる．バーストの度合い $blv(c(h,t))$ は以下の式 (4) のように定義される．

$$blv(c(h,t)) = \begin{cases} -\log(brt(c(h,t))) & (brt(c(h,t)) \leq \beta) \\ 0 & (brt(c(h,t)) > \beta) \end{cases} \tag{4}$$

バースト度合いはデータ構造とは別に保存する．**図 7** のように，データ構造内の各レベルにおいて異なるバースト度合いが算出されるが，期間が重複している部分に関しては，最も大きい値を採用する．各期間 t に対して，最も大きい値を持つバースト度合いを $blv_m(t)$ と表現する．

3.3.3　バーストの重さ

解析期間の各データストリームのバーストをランキング

するための指標として，バーストの重さを定義する．バーストの重さは，バーストしている期間のバースト度合いが正常範囲と比較して，どの程度の差があるかを計算し，期間 W_{min} ごとに合計したものとする．blv_i が各期間のバーストの度合いを表すものとし，各時刻 t に対応する時刻 p_{start} から p_{end} までの期間のバーストの重さ $weight(p_{start}, p_{end})$ を式 (5) のように表現する．

$$weight(p_{start}, p_{end}) = \sum_{i=p_{start}}^{p_{end}} (blv_m(i) - (-\log(\beta))) \quad (blv_m(i) > 0) \quad (5)$$

3.4　バースト解析における数値の補正方法

バースト解析の結果は，解析するデータの種類や監視状況に応じて算出された数値の補正が必要な場合がある．ここでは 3 つの閾値を用いて，少ない発生間隔から過剰なバースト判定を避けるための方法と，長期間にわたるバーストを抑制する方法について説明する．

3.4.1　計算対象のセルに対する過剰なバースト判定を防ぐための処理

バーストの監視対象のイベントについて，突発的に少ない数のイベントが連続して発生した場合（たとえば 2 つのイベントが同時に発生したとき），条件式 (3) を簡単に満たすため，過剰にバーストと判定される．そのため，本手法ではイベント発生間隔の最低個数 A_{min} を設定する．解析対象のセル $c(h,t)$ が，$c(h,t).gapn < A_{min}$ を満たすときは $c(h,t)$ のバースト判定を行わない．このようにして，過剰なバースト判定を避ける．

3.4.2　比較対象のセルに対する過剰なバースト判定を防ぐための処理

バースト監視の初期段階において，新たに生成した $c(h,t)$ に対して $c(N-1, t-1-h)$ が存在しない状況がある．そのため，期間が重複していないセルの中で最新かつ最もレベルが高いセルと比較してバーストを解析する．しかし，図 6 の $c(1,1)$ のように，比較対象のセル内の情報が極端に少ない場合，$c(0,2)$ が過剰にバーストと判定されることがある．そこで，信頼可能な到着間隔の数 C_{min} を設定する．

$tgcell.gapn < C_{min}$ のとき，比較対象のセルの情報の信頼度が $tgcell.gapn/C_{min}$ であると見なして，バースト判定基準に対して制約をかける．$tgcell.gapn < C_{min}$ のとき，条件式 (6) を使用してバースト判定を行う．

$$\frac{avg(c(h,t))}{avg(tgcell)} \times \left(\frac{C_{min}}{tgcell.gapn}\right) \leq \beta \quad (6)$$

このとき，バーストの度合いは次のように評価される．

$$blv'(c(h,t)) = -\log\left(\left(\frac{avg(c(h,t))}{avg(tgcell)}\right) \times \left(\frac{C_{min}}{tgcell.gapn}\right)\right) \quad (7)$$

一般的に，$N > C_{min}$ となるように設定する．

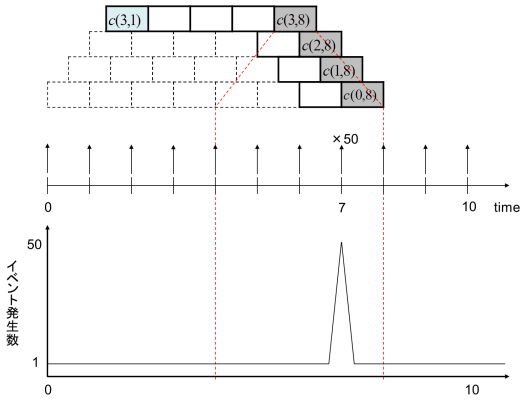

図 8　強いバーストによる周辺期間への影響例
Fig. 8　Example of range of affected cells by high level burst.

3.4.3　長期間にわたるバーストを抑制する方法

バーストを解析していると，**図 8** のようなイベントが突発的に集中して発生する場合がある．このとき $c(3,8)$ に関連する期間もバースト判定されるが，あらかじめ解析データの特徴が分かっており，ある強いバーストの影響による長期間のバーストを抑制する場合，優先させるウィンドウサイズの最大値 W_{max} を設定する．$c(h,t).gaps > W_{max}$ のとき，式 (8) によりバースト判定を行う．

$$\frac{avg(c(h,t))}{avg(tgcell)} \times \left(\frac{c(h,t).gaps}{W_{max}}\right) \leq \beta \quad (8)$$

このとき，バーストの度合いは次のように評価される．

$$blv''(c(h,t)) = -\log\left(\left(\frac{avg(c(h,t))}{avg(tgcell)}\right) \times \left(\frac{c(h,t).gaps}{W_{max}}\right)\right) \quad (9)$$

一般的に，W_{max} は解析に用いる単位時間以上となる．

3.5　解析手順

データ構造内に十分なデータが蓄積され，$c(N-1, t-1-h)$ が存在する場合のバースト解析アルゴリズムを Algorithm 1 に示す．

変数 t はレベル 0 のセルが何回目に生成されたかを表すため，レベル 0 の 1 つのセルに対応するデータを得るたびに Algorithm 1 の 2 行目から実行されることになる．生成されたセル $c(h,t)$ は，条件式 (3) を満たしているかどうか判定される．

バーストを監視し始めた初期段階において，新たに生成した $c(h,t)$ に対して $c(N-1, t-1-h)$ が存在しない場合は，期間が重複していないセルの中で最新かつ最もレベルが高いセルと比較してバーストを解析し，今後使用することがないセルから順次破棄していく．

4.　実験

本実験では，提案手法が新たなリアルタイムバースト解析手法として有用であるかを検証するため，Kleinberg の

Algorithm 1 提案手法のバースト解析アルゴリズム

```
for t do
  h = 0
  while h < N do
    c(h,t) を生成
    c(h−1,t−1) を破棄
    blv(c(h,t)) = 0
    if c(h,t).gapn >= A_min then
      brt(c(h,t)) = avg(c(h,t)) / avg(c(N−1,t−1−h))
      if c(N−1,t−1−h).gapn < C_min then
        brt(c(h,t)) = brt(c(h,t)) * C_min / c(N−1,t−1−h).gapn
      end if
      if c(h,t).gaps > W_max then
        brt(c(h,t)) = brt(c(h,t)) * c(h,t).gaps / W_max
      end if
      if brt(c(h,t)) ≤ β then
        c(h,t) をバーストと判定
        blv(c(h,t)) = −log(brt(c(h,t)))
      end if
    end if
    保持しているバーストの度合いと重さの情報を更新
    if h = N − 1 then
      c(N−1,t−1−h) を破棄
    end if
    c(h,t) をデータ構造に追加
    h++
  end while
end for
```

表 1　実験環境
Table 1　Experiment environment.

OS	Ubuntu11.04
開発言語	C++
CPU	Intel Core i7 2700K @3.50 GHz
メモリ	4 GB

手法 [1] との比較実験を行う．本実験では，監視イベントを単語の出現の有無とする．

4.1　実験概要

実験環境を**表 1** に示す．本実験では，提案手法が現実世界での単語の出現傾向に基づいた監視において有効であるかを検証するため，新聞記事データを利用する．新聞記事データは「CD-毎日新聞データ集 2006 年版」より，あらかじめ加工されて登録されている，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．なお，本実験では「○○日」など日付を表すワードは大半の記事に含まれているため，ノイズとして除外して解析する．

4.2　バースト解析における補正の有効性の評価実験

本実験では，バースト解析におけるパラメータ C_{min}，W_{max} を用いた補正が有効であるか評価するために，パラメータを変化させてバーストを解析した結果を比較する．**図 9** は「会見」が含まれる 1 日あたりの記事数の推移である．基本的なパラメータを $N = 50$，$\beta = 0.4$，$W_{min} = 1$ 日，$A_{min} = 15$，$W_{max} = 1$ 日とし，**図 10** はパラメータ C_{min} を設定せずにキーワード「会見」について解析した結果，**図 11** は $C_{min} = 15$ と設定して解析した結果である．両者の違いは経過日数 5 付近でのバーストの有無である．C_{min} を設定しない場合，監視開始直後で解析に必要な情報が不足している状況であるため，イベント発生数があまり高くなくても，バーストとして検出される．一方，C_{min} を設定した場合，監視直後のイベント数が少ない場合の過剰な反応を抑えている．このことから，監視を開始して情報が少ない場合に，C_{min} が有効であることが分かる．

図 12 は「トーナメント進出」が含まれる 1 日あたりの記事数の推移である．基本的なパラメータを $N = 50$，$\beta = 0.4$，$W_{min} = 1$ 日，$A_{min} = 15$，$C_{min} = 15$ と設定し，**図 13** はパラメータ W_{max} を設定せずに解析した結

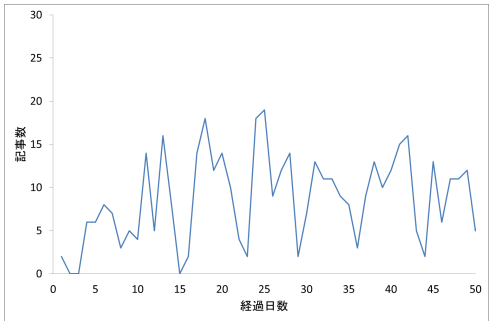

図 9　「会見」が含まれる記事数
Fig. 9　Number of articles including “interview”.

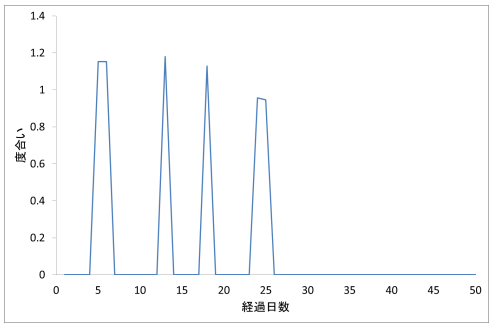

図 10　C_{min} を設定しない場合の提案手法による「会見」の解析結果
Fig. 10　Result of analysis of “interview” by proposed method without C_{min}.

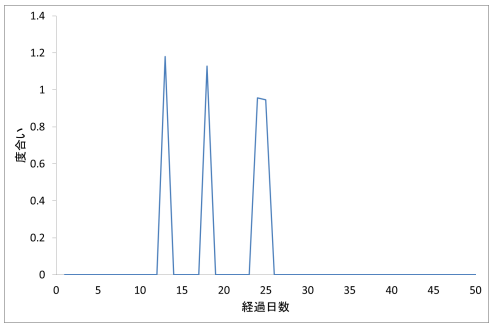

図 11　提案手法による「会見」の解析結果
Fig. 11　Result of analysis of “interview” by proposed method.

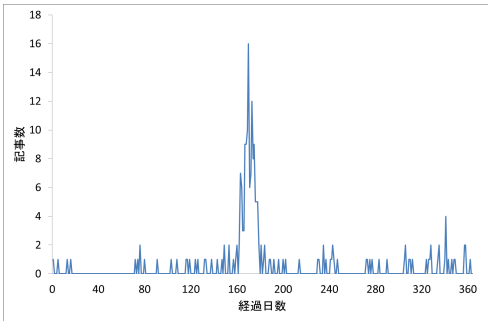

図 12　「トーナメント進出」が含まれる記事数
Fig. 12　Number of articles including “advance to tournament”.

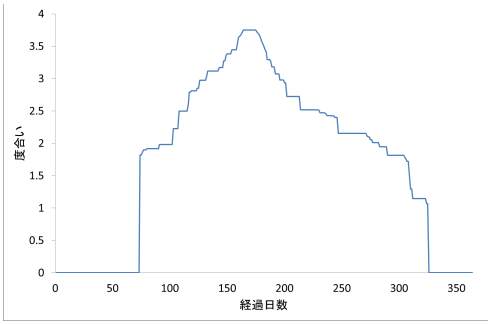

図 13　W_{max} を設定しない場合の提案手法による「トーナメント進出」の解析結果
Fig. 13　Result of analysis of “advance to tournament” by proposed method without W_{max}.

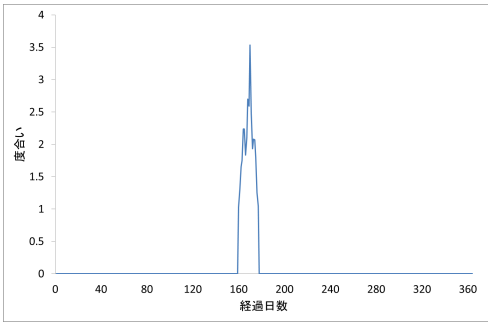

図 14　提案手法による「トーナメント進出」の解析結果
Fig. 14　Result of analysis of “advance to tournament” by proposed method.

果，**図 14** は $W_{max} = 1$ 日と設定して解析した結果である．図 13 はイベント発生数が多い期間を中心に，広範囲にバーストが検出されている．これは発生数が多い期間が含まれる高レベルのセルの到着間隔の平均値が短くなり，そのセルの期間すべてがバーストと判定されているためである．一方，W_{max} を設定した場合，検出されるのは短期間のバーストのみであり，新聞記事のバーストを 1 日ごとに表示する場合に，実際の出現数と似た形でバーストが検出される．このことから，短期間のバーストが長期間にわたって与える影響を抑える場合に，W_{max} は有効であることが分かる．

4.3　バースト検出精度の比較実験

本実験では，新聞記事データの記事から得られた本文キーワードの出現頻度上位 10,000 件を対象として，提案手法によりバースト解析した結果と Kleinberg の手法によりバースト解析した結果とを比較する．本評価では，次に示す 3 つの評価を実施する．

- 各手法で得られたバーストの重さによるランキングの上位 80 単語の比較
- 各手法で得られたバーストの重さによるランキングの一致率の評価
- 各手法で得られたバーストの度合いのグラフの特性の比較

すべての実験において提案手法のパラメータは予備実験の結果に基づき，経験的に $N = 50$，$\beta = 0.4$，$W_{min} = 1$ 日，$A_{min} = 15$，$C_{min} = 15$，$W_{max} = 1$ 日を採用した．

4.3.1　バーストの重さによるランキング上位 80 単語の比較結果

各手法のバースト解析結果より得られたバーストの重さランキング上位 80 単語を**表 2**，**表 3** に示す．これらの表

表 2　提案手法によるバーストワードランキング
Table 2　Burst ranking by proposed method.

順位	単語	順位	単語
1	安倍晋三首相	41	飲酒運転
2	パ交流戦	42	クロアチア戦
3	パ交流	43	交流戦
4	履修	44	１５ｔｈ
5	ミサイル発射	45	ＭＡＮＹ
6	Ｇａｍｅｓ	46	ＧＥＲＭＡＮＹ
7	不足問題	47	ＧＥＲＭＡ
8	核実験	48	ＧＥＲ
9	実験実施	49	ＥＲＭＡＮＹ
10	核実験実施	50	ＥＲＭＡ
11	ヒズボラ	51	酒気帯び
12	安倍首相	52	前知事
13	レバノン	53	発射
14	制裁決議	54	木村良樹
15	良樹	55	入札制度
16	愛媛県宇和島市	56	米中間選挙
17	やらせ	57	間選挙
18	ジャワ島	58	塩崎恭久
19	組織ヒズボラ	59	佐藤栄佐久
20	レバノン南部	60	オーストラリア戦
21	宇和島	61	酒気帯び運転
22	宇和島市	62	ｔｏｒｉｎｏ２００６
23	徳洲会病院	63	ミサイル
24	首相補佐官	64	ＡＰＥＣ
25	徳洲会	65	弾道ミサイル
26	セ・パ交流戦	66	太平洋経済協力会議
27	セ・パ交流	67	太平洋経済協力
28	核保有	68	アジア太平洋経済協力会議
29	出納長	69	アジア太平洋経済協力
30	シーア派民兵組織	70	太平洋経済
31	イスラム教シーア派民兵組織	71	ＥＲ
32	イスラム教シーア派民兵	72	アジア太平洋経済
33	タウンミーティング	73	世界ランク
34	シーア派民兵	74	ＤＯＨＡ
35	セ・パ	75	腎臓
36	北越	76	コスタリカ
37	決勝トーナメント進出	77	実弟
38	拿捕	78	ＮＹ
39	Ａｓｉａｎ	79	米中間
40	トーナメント進出	80	インサイダー

表 3　Kleinberg の手法によるバーストワードランキング
Table 3　Burst ranking by Kleinberg's method.

順位	単語	順位	単語
1	安倍晋三首相	41	スキー
2	ＭＡＮＹ	42	１回戦
3	ＧＥＲＭＡＮＹ	43	冬季
4	ＥＲＭＡＮＹ	44	ブラジル
5	ＥＲＭＡ	45	女子
6	ＧＥＲＭＡ	46	交流戦
7	ＧＥＲ	47	総裁選
8	ＥＲ	48	ドイツ大会
9	ＮＹ	49	ミサイル発射
10	W杯	50	甲子園
11	ＭＡ	51	決議
12	パ交流戦	52	セ・パ
13	パ交流	53	セ・パ交流戦
14	ｔｏｒｉｎｏ２００６	54	セ・パ交流
15	１次	55	安倍晋三官房長官
16	トリノ	56	準決勝
17	１次リーグ	57	登板
18	Ｇａｍｅｓ	58	ヒズボラ
19	Ａｓｉａｎ	59	堀江
20	１５ｔｈ	60	決勝トーナメント
21	ＤＯＨＡ	61	村上ファンド
22	ライブドア	62	知事
23	五輪	63	トーナメント
24	核実験	64	２回戦
25	北朝鮮	65	金メダル
26	男子	66	談合
27	レバノン	67	負
28	安倍	68	冬季五輪
29	実験	69	ファンド
30	安倍晋三	70	交流
31	ミサイル	71	スケート
32	サッカー	72	村上
33	いじめ	73	今大会
34	決勝	74	６カ国協議
35	メダル	75	ホームチーム
36	核	76	安倍氏
37	ドイツ	77	安倍首相
38	制裁	78	イスラエル
39	リーグ	79	安保理
40	発射	80	現地

を確認すると，次に示す 3 つのことが明らかとなった．

1 つ目として，提案手法では，上位 80 以内にランクインしている単語は，継続的に出現する単語ではなく，「履修」，「ミサイル発射」，「飲酒運転」，「ＡＰＥＣ」，「シーア派民兵組織」など，時事的な話題に関するものが多く，突発的に話題としてのぼるデータが上位にくる傾向があることが分かった．そのため，ランクインしている単語が関連するトピックも散在しており，突発的に発生する単語を多く抽出できていることが明らかとなった．これより，提案手法では，つねに頻出する単語が上位にランクインする出現頻度のランキングとは異なる性質の結果を得られることが分かる．

2 つ目として，Kleinberg の手法では，上位 80 以内にランクインしている単語は，同一の話題として，「サッカーワールドカップドイツ大会」，「トリノオリンピック」，「核実験」に関連するキーワードが多く見られた．この結果を分析すると，Kleinberg の手法では，突発的に発生するキーワードよりも，社会の注目が高まっており，一定期間以上継続して話題になるキーワードを抽出できることが分かった．

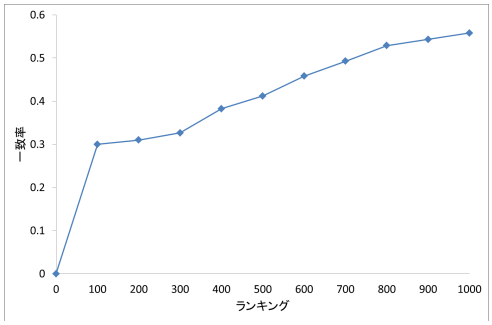

図 15　提案手法と Kleinberg の手法のバーストワードランキング上位ワードの一致率
Fig. 15　Agreement rate of burst ranking between Kleinberg's method and proposed method.

3 つ目として，提案手法と Kleinberg の手法の解析結果を比較すると，提案手法で 40 位以内のキーワードの多くが Kleinberg では 40 以下となり，一方，提案手法で 40 位以下のキーワードの多くが Kleinberg では 40 以内に出現するという状況となっており，ランキングに偏りが見られることが分かった．これは，提案手法において，長期間，同一の話題に関する記事が多く出現している場合は，バースト状態が継続的に続いているのではなく，バースト状態が収束傾向に向かっていると判断するためであると考えられる．

上述の 3 つの点を考慮すると，提案手法と Kleinberg の手法で得られるバーストの重さは異なるキーワードを分析して抽出できることが分かる．具体的に，提案手法では，突発的に発生する時事的なイベントに非常に強く，イベントが発生した直後の時点で，鋭敏に反応して出力できていることが分かる．そのため，リアルタイムに新たに発生した注目の話題などを即座に発見して取得することに役立てることができると考えられる．一方，Kleinberg の手法では，社会的に一定期間継続して話題となるキーワードを抽出できる．そのため，後々の各年度の主要トピックをまとめるなどの際に，有効に活用することができると考えられる．これらのことから，提案手法と Kleinberg の手法の解析結果を併用することで，バーストの特性を考慮した幅広いアルゴリズムやシステムを構築することができると考えられる．

4.3.2　バーストの重さによるランキングの一致率の評価

各手法のバースト解析結果の重さによるランキングの上位キーワードの一致率を**図 15** に示す．一致率のグラフを確認すると，上位 100 件で約 3 割，500 件で約 4 割となっており，一致率が低いことが分かる．各手法の詳細を確認すると，Kleinberg の手法では，同様のイベントに関わるキーワードが多く含まれており，キーワードからトピックを一目で推定できるような社会的に注目されているものが

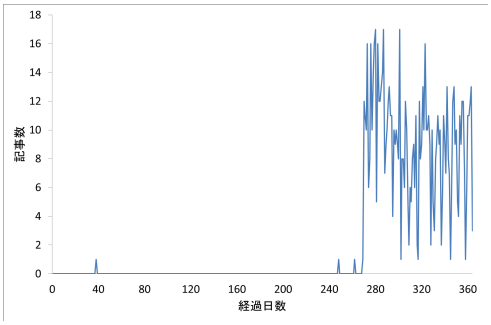

図 16 「安部晋三首相」が含まれる記事数
Fig. 16 Number of articles including “Prime Minister Abe Shinzo”.

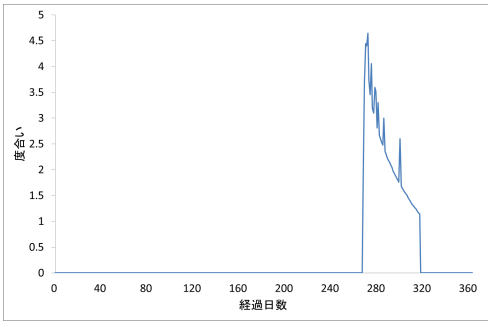

図 17 提案手法による「安部晋三首相」の解析結果
Fig. 17 Result of analysis of “Prime Minister Abe Shinzo” by proposed method.

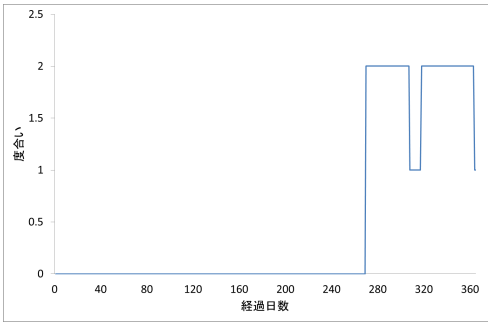

図 18 Kleinberg の手法による「安部晋三首相」の解析結果
Fig. 18 Result of analysis of “Prime Minister Abe Shinzo” by Kleinberg’s method.

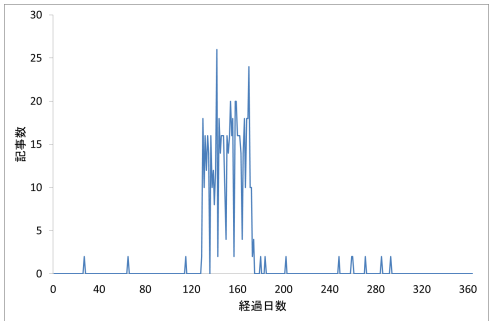

図 19 「パ交流戦」が含まれる記事数
Fig. 19 Number of articles including “Pa interleague game”.

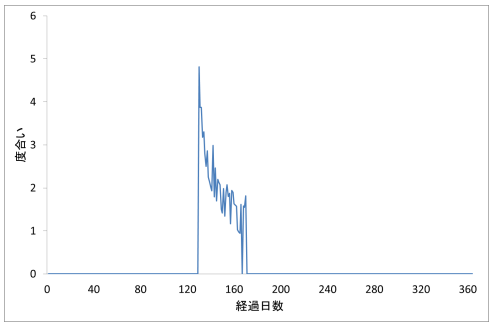

図 20 提案手法による「パ交流戦」の解析結果
Fig. 20 Result of analysis of “Pa interleague game” by proposed method.

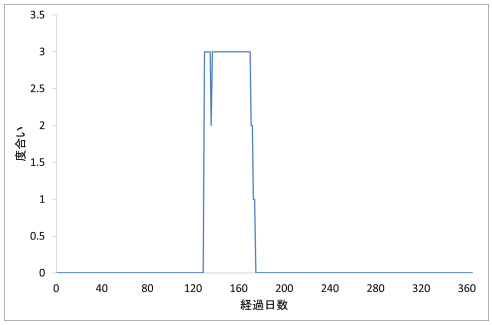

図 21 Kleinberg の手法による「パ交流戦」の解析結果
Fig. 21 Result of analysis of “Pa interleague game” by Kleinberg’s method.

多く見られた．それに対して，提案手法では，キーワードの種類が多岐にわたり，様々なトピックから抽出されている状況であった．これらのことから，提案手法は意外性が高く，突発的に集中発生するイベントの抽出に効果が高いことが明らかとなった．

4.3.3 バーストの度合いのグラフの特性の比較

提案手法のバーストの重さランキング上位 80 単語を示した表 2 のうち，Kleinberg の手法の抽出結果と一致した，上位 4 件の単語に対して，バーストの度合いの比較を行う．バーストの度合いの比較では，それぞれの単語が含まれる記事数，提案手法と Kleinberg の手法で求めたバーストの度合いのグラフの 3 つを用いて，それぞれの手法のバーストの度合いの特性を評価する．上位 4 件の単語「安倍晋三首相（**図 16**，**図 17**，**図 18**）」，「パ交流戦（**図 19**，**図 20**，**図 21**）」，「ミサイル発射（**図 22**，**図 23**，**図 24**）」，「Ｇａｍｅｓ（**図 25**，**図 26**，**図 27**）」の評価結果を確認すると，次の内容が明らかとなった．

安倍晋三首相の記事数の解析結果（図 16）を確認すると，経過日数約 270–365 日までつねに記事が存在しており，定期的に継続して話題が発生していることが分かる．図 18 を確認すると，Kleinberg の手法ではバーストの度合いはつねに高い値を示していることが分かる．それに対して，図 17 を確認すると，提案手法では一定時間が経過したのちにバーストの度合いが減少しており，320 日の段階で 0 となっていることが分かる．また，パ交流戦（図 19–21）

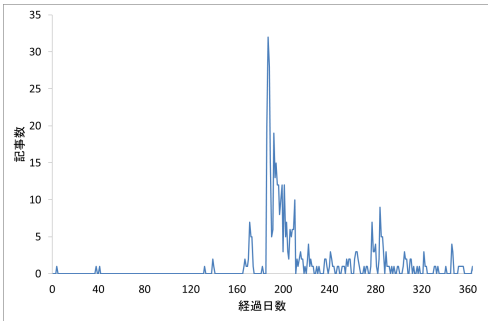

図 22　「ミサイル発射」が含まれる記事数
Fig. 22　Number of articles including “missile launch”.

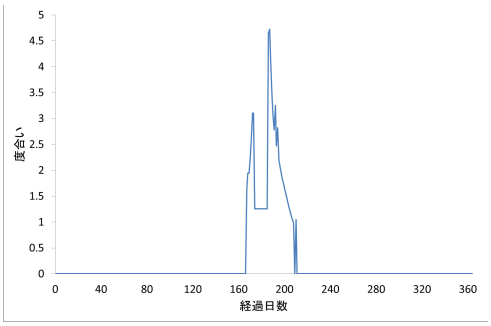

図 23　提案手法による「ミサイル発射」の解析結果
Fig. 23　Result of analysis of “missile launch” by proposed method.

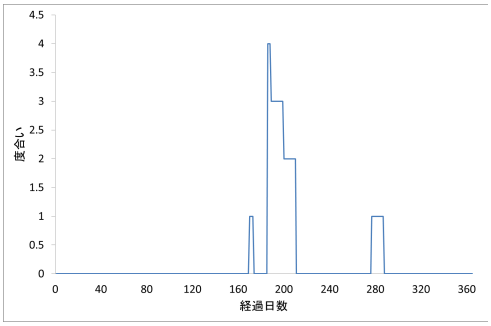

図 24　Kleinberg の手法による「ミサイル発射」の解析結果
Fig. 24　Result of analysis of “missile launch” by Kleinberg’s method.

においても同様の結果が見られる．このことより，提案手法は発生数が同じ期間内でも，急上昇している部分をバーストの度合いが高いと評価していることが分かる．この特性は，つねに最新の注目トピックを判定して抽出できるという点で，リアルタイムバースト解析を目的とした場合は，非常に有用であると考えられる．

なお，リアルタイム性の評価については文献 [5] を参照されたい．

5.　おわりに

本論文では，データストリーム中のバーストの度合いやランキングをリアルタイムに解析する手法を提案した．これは，一定期間ごとではなくイベント発生時にバーストの

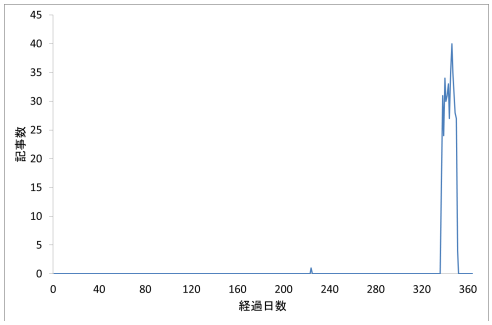

図 25　「Ｇａｍｅｓ」が含まれる記事数
Fig. 25　Number of articles including “Games”.

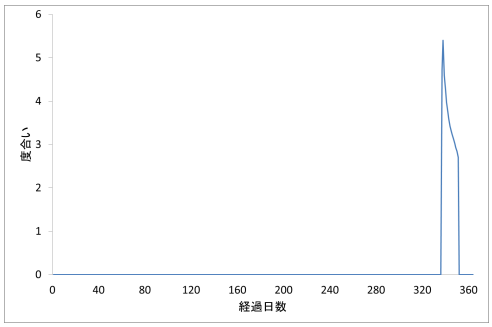

図 26　提案手法による「Ｇａｍｅｓ」の解析結果
Fig. 26　Result of analysis of “Games” by proposed method.

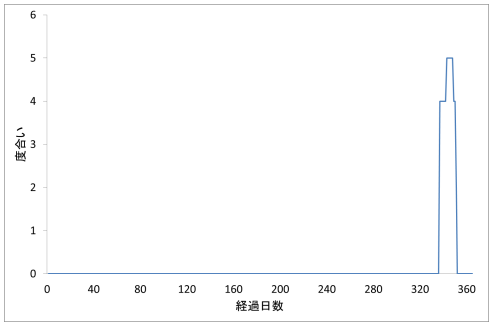

図 27　Kleinberg の手法による「Ｇａｍｅｓ」の解析結果
Fig. 27　Result of analysis of “Games” by Kleinberg’s method.

解析を行うことにより，イベントが発生していない無駄なデータ更新を防ぐ．また，イベントが集中発生している期間のデータを圧縮して保持するため，膨大な量のイベントが集中発生したときに，データの更新を一定に制限できる．さらに，バーストの度合いやランキングを算出することにより，これらに基づいた効率的な解析が可能となる．

実験により，提案手法は社会的に注目の高い一定期間発生するイベントよりも，短期間で突発的に発生するイベントを選択的に抽出可能であることが確認できた．これは，解析した一時点において，急上昇したキーワードを抽出することが可能であることから，リアルタイム解析において目新しい意外性の高いものを選出できるため，日々時々刻々と変化するイベントをリアルタイムに監視する場合には有用であると考えられる．

本提案手法の応用として，突発的に短期間で発生するイ

ベントをリアルタイムに解析できるという特徴を有するため，サーバのアクセスログの解析などに利用することで，短期的な DOS 攻撃のログを選択的に発見可能であると考えられる．

今後の課題として，新聞記事以外の，時系列に従ったイベントの情報に対しても提案手法を適用し，提案手法の応用可能性について検証したいと考えている．

参考文献

[1] Kleinberg, J.: Bursty and Hierarchical Structure in Streams, *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.91–101, ACM (2002).

[2] Zhu, Y. and Shasha, D.: Efficient Elastic Burst Detection in Data Streams, *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.336–345, ACM (2003).

[3] Shasha, D. and Zhu, Y.: *High Performance Discovery in Time Series: Techniques and Case Studies* (*Monographs in Computer Science*), Springer-Verlag (2004).

[4] Zhang, X. and Shasha, D.: Better Burst Detection, *Proc. 22nd International Conference on Data Engineering*, pp.146–149, IEEE Computer Society (2006).

[5] 蝦名亮平，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, pp.1–6 (2010).

[6] Kumar, R., Novak, J., Raghavan, P. and Tomkins, A.: On the Bursty Evolution of Blogspace, *Proc. 12th International Conference on World Wide Web*, pp.568–576, ACM (2003).

[7] Platakis, M., Kotsakos, D. and Gunopulos, D.: Discovering Hot Topics in the Blogosphere, *Proc. 2nd Panhellenic Scientific Student Conference on Informatics, Related Technologies and Applications*, pp.122–132 (2008).

[8] He, Q., Chang, K. and Lim, E.-P.: Using Burstiness to Improve Clustering of Topics in News Streams, *Proc. 2007 7th IEEE International Conference on Data Mining*, pp.493–498, IEEE Computer Society (2007).

[9] He, Q., Chang, K., Lim, E.-P. and Zhang, J.: Bursty Feature Representation for Clustering Text Streams, *Proc. 7th SIAM International Conference on Data Mining*, pp.491–496, SIAM (2007).

[10] Lappas, T., Arai, B., Platakis, M., Kotsakos, D. and Gunopulos, D.: On Burstiness-Aware Search for Document Sequences, *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.477–486, ACM (2009).

[11] Sakkopoulos, E., Antoniou, D., Adamopoulou, P., Tsirakis, N. and Tsakalidis, A.K.: A Web Personalizing Technique Using Adaptive Data Structures: The Case of Bursts in Web Visits, *Journal of Systems and Software*, Vol.83, pp.2200–2210 (2010).

蝦名 亮平　（学生会員）

1987 年生．2010 年立命館大学情報理工学部卒業．2012 年同大学大学院理工学研究科博士前期課程修了．2012 年同大学院情報理工学研究科博士後期課程入学，現在に至る．データマイニングの研究に従事．日本データベース学会学生会員．

中村 健二　（正会員）

1981 年生．2004 年関西大学総合情報学部卒業．2006 年同大学大学院総合情報学研究科知識情報学専攻博士課程前期課程修了．2009 年同研究科総合情報学専攻博士課程後期課程修了．2009 年関西大学ポスト・ドクトラル・フェロー．2010 年立命館大学情報理工学部情報システム学科助手．2012 年大阪経済大学情報社会学部准教授，現在に至る．博士（情報学）．知識情報処理，テキストマイニング，Web マイニング等の研究に従事．2002～2012 年（株）関西総合情報研究所にて活動．システム設計，データモデル設計等の研究開発に従事．土木学会，日本データベース学会各会員．

小柳 滋　（正会員）

1949 年生．1977 年京都大学大学院博士課程修了．1977 年（株）東芝入社．2002 年立命館大学教授，現在に至る．博士（工学）．データマイニング，コンピュータアーキテクチャの研究に従事．IEEE-CS，ACM，電子情報通信学会，日本データベース学会各会員．

（担当編集委員　上善 恒雄）