社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－NL－160 (13)
2004／3／5

document streamにおけるburstの発見

藤木 稔明† 南野 朋之† 鈴木 泰裕† 奥村 学‡

概 要

近年、Web上には様々な意見情報が存在している。そのため、Web上でいつ、どのような情報が注目されていたかを把握することは企業活動にとって重要な情報になりつつあると考えられている。本研究では特にblogと電子掲示板を対象とし、その中で注目されている話題を示すトピックワードを発見するための手法を提案する。提案手法は、Kleinbergの提案するburst検出手法に基づいた手法であり、blogや電子掲示板に対する書き込みを時間情報を含む文書列であるdocument streamとして扱い、その中でdocumentの出現間隔が短くなっている箇所(burst)に注目すべきイベントがあるとして検出する。その際、Kleinbergの提案する手法ではイベント発生と無関係にdocument数が変動するようなdocument streamをうまく扱うことができないため、手法の拡張を行った。また提案手法を用いてトピックワードの抽出実験を行い、手法の有効性を評価した。

Identification of bursts in a document stream

Toshiaki FUJIKI† Tomoyuki NANNO† Yasuhiro SUZUKI† Manabu OKUMURA‡

Abstract

The reputation is now disseminated quickly on the WWW, because everyone can send a message to the world easily by using blog or BBS. Therefore, it is highly required to find out what information attracts people's attention and what opinion they have. We propose a method for extracting 'burst of a word' which is related to a popular topic in a document stream. A document stream is defined as a sequence of documents which arrive in temporal order, and we regard blog and BBS as document streams to apply the method originally proposed by Kleinberg. However, since Kleinberg's algorithm cannot be applied to the document streams whose distribution of documents is not uniform, we extend the method to be able to apply to blog and BBS. We also describe experiments for blog and BBS with our proposed method and discuss the results.

1 はじめに

インターネットの普及と共にWeb上で個人が情報を発信する機会が多くなっている。個人の発信する情報としては、電子掲示板やメーリングリスト、日記や個人ニュースサイトの総称であるblogなどが挙げられるが、これらは自分の身近に存在する情報や興味のある情報、経験したこと、感じたことをそのまま不特定多数の読者に対して発信するという特徴を持っている。これらの情報源は、同じ興味や考えを持つ人々にとって重要な情報源となっているだけでなく、少ないコストで直接消費者の声を収集できるという意味で、企業からも注目され始めている。

しかし逆に、情報量が膨大であるため人手で監視することの出来る範囲には限界があり、このような個人発信の情報源全体を考えたときにはどのような話題が注目されているのかを把握することも難しい。そのため、監視していない情報源で大きな事件が起きていてもそれを知ることが出来ないということが起きる可能性もある。このような問題に対処するため、我々はblogの自動収集と監視を行うシステムを試作している[1]。このシステムは特定のツールが生成する書式やメタデータに依存することなく、HTML文書を解析することによってblogであるかどうかを判断し、網羅的に収集、監視することを目的としており、高い精度で大量のblogを収集することが可能であることを示している。また、このシステムは収集されたblogを検索する機能を持っており、利用者が任意の検索語を入力することで、その語を含むblog記事(以下、blogを一日単位に分割した際の項目を指すとする)を表示することが可能となっている。しかし、利用者が"去年の6月にあった出来事を知りたい"などといった抽象的な検索要求しか持たない場合、通常のキーワードによる検索を行うことはできない。このように、明確な検索要求をもてない利用者に対してはシステム側があらかじめキーワード候補を用意し、提示する仕組みが必要となる。

一方、blogや電子掲示板に対する書き込みなどでは、ある話題に注目が集まる場合に特定の語句の出現頻度

†東京工業大学大学院 総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
{fujiki,nanno,yasu}@lr.pi.titech.ac.jp
‡東京工業大学 精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology
oku@pi.titech.ac.jp

が急激に上がるといった現象が起こる。これは、多くの人がその話題に注目し、言及することによって、その話題に関連する固有名詞などが一時的に出現しやすくなるためである。

そこで本研究では前述のblogと電子掲示板「2ちゃんねる:ニュース速報+」を対象とし、このような一時的に出現頻度の上昇した単語をその時期におけるトピックワードとして抽出する手法を提案する。これは、トピックワードを抽出することによってその情報源ではどのような話題について盛り上がっているのかを知ることができると考えるためである。そのために本研究ではKleinbergの提案するburst検出手法[2]に注目する。この手法はblogや電子掲示板への書き込みを、新聞記事などのように時間情報のついた文書の集合を意味するdocument streamとしてとらえ、document stream中でdocument数が急激に増加している部分(burst)を発見する手法である。ただし後述のように、本研究が対象とするようなdocument streamに対しては単純に適用することができない。よって、本研究ではこの手法を拡張した手法を提案し、その手法を用いてblogと電子掲示板からのトピックワード抽出実験を行い、その結果を評価する。

2 関連研究

電子掲示板を対象としてトピックワードを発見することを目的とした研究として松村らの研究[3]がある。この研究では電子掲示板上の書き込みに含まれる単語が、その書き込みへの返信でも使われるということを単語の伝搬と考える。そして、このような単語は話題を示す中心的な単語であり、伝搬の長さは話題の長さを示すとされている。ただしこの研究では、Yahoo!掲示板[1]のように書き込み間に親子関係が明示され(ある書き込みに対して返信する形で書き込む)、木構造となるような電子掲示板のみを対象としている。このような書き込み間の関係を利用することによって、並列に複数の話題が存在するような場合にも話題を分離して解析することが可能となるが、逆に、処理対象とすることができる電子掲示板は限られてしまう。

時系列に並んだ文書集合からのトピックワード抽出に関する研究としては、ある期間の新聞記事に含まれるイベントを発見することを目的とする白倉らの研究[4]が挙げられる。この研究ではトピックワード候補として名詞と動詞を考え、それらがある数日間に集中して現れるような場合にその語をトピックワードとしている。このとき、集中していることを示す尺度としては、(その期間における出現頻度の合計)/(全期間の出現頻度の合計)という値を利用している。そのため、総出現頻度が低く、短い期間に出現するような語を扱うための足きり条件が必要であるほか、話題の続く長さをあらかじめ仮定する必要があるとされている。

仲村ら[5]は、新聞記事中に出現する固有名詞の出現頻度が異常に高くなっている場合をイベントの発生とみなすことで、イベントの検出を行っている。この手法ではそれぞれの固有名詞はある確率分布に従っていると仮定されており、その確率分布において閾値として危険率を設定、そこから導かれる回数以上に固有名詞が出現している場合に異常出現であるとみなしている。そして、異常出現の日が連続している場合に、その固有名詞に関するイベントがその期間で起こっていると考えられるとしている。この研究は単語の出現頻度が確率分布に従うとする点でKleinbergの手法や提案手法と類似している。しかし、提案手法ではイベントが発生している状態に対して異なる確率分布を定義し、どちらの分布がうまく近似しているかに基づいて判断しているという点で異なっている。また、提案手法ではこのように2つの確率分布を使い分けることによって、状態遷移といった概念を用いることが可能となっている。

[1] http://messages.yahoo.co.jp/index.html

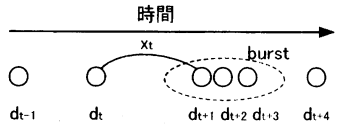

図 1: document stream

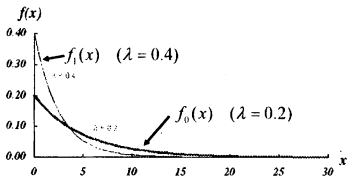

図 2: 指数分布

3 burstの検出

本研究ではKleinbergの提案するburst検出手法[2]を拡張した手法を用いて、blogと電子掲示板におけるトピックワードの検出を行う。そこでまずKleinbergの提案する手法について説明した後、提案手法について説明する。

3.1 Kleinbergの提案するburst検出手法

この手法はdocument streamにおけるburstを検出する手法である。ここでdocument streamとはdocument d_tが到着した後、x_tの間隔をおいて次のdocument d_{t+1}が到着するというようなdocument集合からなるstreamのことを指す(図1)。例えば電子メールの受信を考えると、あるメールd_tが到着した後、x_t時間後に次のメールd_{t+1}が到着するというdocument streamのモデルで表現することができる。同様に、新聞記事や電子掲示板、blogなどもdocument streamと考えることができる。

document streamにおいては、イベントの発生に関連してdocument数が増減し、x_tが変化することが仮定できる。例えば、あるニュースに関連する新聞記事の記事数は、そのニュースが発生した直後に増え、その後一時減少しても、関連する事件が起きた際に再び増加するという増減を繰り返していると考えられる。

このような場合、記事の増加している部分はその記事に関連するイベントが発生した時期を示していると考えられ、この期間を平常状態に対して burst 状態と呼ぶ。そしてこのような状態変化は、document 数の変化、つまり、到着間隔 x_t の変化から発見することができる。よって、特定のイベントについての盛り上がり部分 (burst) を発見したい場合は、そのイベントに関連する document だけを元の document stream から抜き出し、その document stream 中で document の到着間隔が短くなっている部分を発見すればよい。

Kleinberg の提案する手法では、document の到着間隔 x_t は指数分布に従うと仮定して計算を行う。これは document がランダムに到着すると仮定することと同義であるためだと考えられる。すなわち、ある時間区間 $(0,T]$ に K 個の document が到着し、その到着確率が一様分布であるとき、T 中の幅 x の区間 $(y,y+x]$ にある document の到着時間 τ が含まれる確率は

$$P(\tau \in (y,y+x] \subset (0,T]) = x/T \tag{1}$$

で与えられる。よって、時間区間 $(y,y+x]$ に到着する document 数を $A((y,y+x])$ で表すと、区間内に k 個の document が到着する確率は

$$P(A((y,y+x]) = k) = \frac{K!}{k!(K-k)!}(\frac{x}{T})^k(1-\frac{x}{T})^{K-k} \tag{2}$$

となる。ここで、$\lambda = K/T$ と定義し、$T \to \infty$、$K \to \infty$ を考えると、式 (2) は

$$P(A((y,y+x]) = k) = e^{-\lambda x}\frac{(\lambda x)^k}{k!} \tag{3}$$

と変形できる。これは、ある時間区間内に到着する document 数はポアソン分布に従うことを示している。さらに、時刻 t_0 に document が到着することを $Z(t_0)$ と表し、その条件下で次の document の到着までの間隔 X を考えると、

$$\begin{aligned} P(X > t \mid Z(t_0)) &= P(A((t_0,t_0+t]) = 0 \mid Z(t_0)) \\ &= P(A((t_0,t_0+t]) = 0) \\ &= e^{-\lambda t} \end{aligned} \tag{4}$$

よって、到着間隔の累積分布関数と確率密度関数は以下の指数分布で表されることがわかる。

$$P(X < t \mid Z(t_0)) = 1 - e^{-\lambda t} \tag{5}$$

$$P(t) = \lambda e^{-\lambda t} \tag{6}$$

つまり、以上を考えると Kleinberg の手法で到着間隔が指数分布に従うと仮定されているのは、document の到着がランダムであると仮定したためであることがわかる。

一方、指数分布とは確率密度関数が $f(x) = \lambda e^{-\lambda x}$ で表される分布であり、グラフにすると図 2 となる。このグラフからわかるように 2 つのグラフ $f_0(x) = \lambda_0 e^{-\lambda_0 x}, f_1(x) = \lambda_1 e^{-\lambda_1 x}$ $(\lambda_0 < \lambda_1)$ に対し、グラフの交点よりも大きな x については $f_0(x) > f_1(x)$ であり、交点よりも小さな x については $f_0(x) < f_1(x)$ である。

ここで指数分布のパラメータ λ は前述のように $\lambda = K/T$、つまり単位時間あたりの document 数と定義される。そこで λ_0 を平常状態に対応する指数分布のパラメータとし、λ_1 を burst 状態での指数分布のパラメータとし、パラメータ s を用いて $\lambda_1 = s\lambda_0$ であるとする。すると、ある x_t が $f_0(x_t) < f_1(x_t)$ を満たすことは λ_1 に由来する確率密度関数の与える値が λ_0 に由来する確率密度関数よりも大きな値となることを意味し、つまり、x_t の時間間隔に対しては burst 状態である可能性が高いということを意味する。よって同様に、ある x_t に対し $f_0(x_t) > f_1(x_t)$ であればその x_t に対しては平常状態である可能性の方が高いことになる。

以上の方法に基づくと、全ての document d_t について、その出現間隔 x_t に対する $f_0(x_t)$ と $f_1(x_t)$ の値を比較することでその document が burst 状態であるか平常状態であるかを判定することができる。しかしこのように単純に判定をしてしまうと、burst 状態と平常状態とが簡単に移り替わってしまい、細かな burst が多くできることになる。これは、ある程度連続する burst はまとめて 1 つの burst であるとして取り扱いたい場合に不利である。そこで、ある document stream の状態列を、通常状態 (状態 0) と burst 状態 (状態 1) の 2 状態からなるオートマトンを用いて表し、全ての document と状態に関して必要なコストを計算した後、最小コストとなるような状態列を計算する手法を用いる。そしてその際に、状態遷移コストを設けることで、平常状態から burst 状態への遷移が簡単に起こらないようにする。

実際には以下のような手順で計算を行う。以下では、document 番号を t、$f_j(x)$ を λ_j をパラメータとする指数分布の確率密度関数、t において状態 j であるために必要なコストを $C_j(t)$、$\tau(l,j)$ を状態遷移コストとそれぞれ表し、状態 0 は平常状態、状態 1 は burst 状態を意味することとする。

1. 初期状態 $t=0$ について、$C_0(t)=0, C_1(t)=\infty$ とする
2. $t = t+1$
3. $C_j(t)$ $(j=0,1)$ を計算する
   ここで
   $$C_j(t) = -\ln f_j(x_t) + \min_l(C_l(t-1)+\tau(l,j)) \tag{7}$$
   とする。ここで j は d_t における状態を表し、l は d_{t-1} における状態を表す。また、状態遷移コスト $\tau(l,j)$ は $l<j$ ならば $\tau(l,j)=\gamma$、それ以外の場合は $\tau(l,j)=0$ とする。
4. 全ての document について 2,3 を繰り返す
5. コストが最小である状態列を選択する
   Viterbi アルゴリズム [6] によって、最後の document d_{last} から逆にたどっていく。つまり、$C_j(last)(j=0,1)$ の小さい j を d_{last} の状態とし、その $C_j(t)$ を与えた前状態 l を d_{t-1} の状態とする。これを

繰り返すことによって、最小コスト状態列を求めることができる。

このようにして各 document に状態を付与した後で、burst 状態である document が連続する部分を一つの burst であるとして取り扱う。すると一つの burst の期間は、その burst に含まれる document 集合の最初の document の時間から最後の document までの時間として定義できる。

またこのとき、ある d_t を平常状態と考えるよりも burst 状態と考えることによって改善されるコストの大きさを d_t が burst 状態である確からしさを示すスコア (burst 度) として用いることができる。この値を用いると、連続して burst 状態になっている document 集合に対する burst 度も計算することができる。本研究では、burst 部分を構成する全ての document に対する burst 度の総和をその burst 部分の burst 度スコアとして用いた。これは予備実験を行った中で、もっともよい結果が得られたためである。

Kleinberg の手法では他の手法に比べて以下の点が優れているとされている。

- 小さな burst を作らず、ある程度大きな単位で扱うことができる
- document stream 毎の平均 document 出現間隔に基づいて計算が行われるため、document の出現頻度が高いものと低いものを同列に扱うことができる
- 比較的計算量が少ない

3.2 Kleinberg の手法の拡張

前節で説明した burst 検出手法は、イベント発生に基づく document 数の増減を検出するための手法である。しかし、この手法を適用するためには document stream の document 数が平常状態において一定である必要がある。これは新聞記事や電子メールなどにおいては成り立つ仮定であるが、本論文が対象とする電子掲示板や blog においては成り立たない。

図 3 は電子掲示板「2 ちゃんねる:ニュース速報+」に対する 5 日間の書き込み数を示したものである。このグラフでは x 軸を時間、y 軸を書き込み数として、20 分間隔でその間の書き込み数をプロットしている。この図を見ると、昼夜で書き込み数の平均値が大きく違うことを読みとることができる。また、我々の収集した blog においても図 4 のように、時期によって総数が大きく異なっている。これはインターネットの普及と共に全体の数が増加しているためであると考えられる。

これらの document stream では、イベント発生以外の要因によって document 数の増減が起こっていると考えられるため、平常状態における document 数が一定であるという仮定をすることができない。そのため、Kleinberg の手法はうまく働かないことが予想される。そこで本節ではそのような影響を受けずに burst 検出ができるように拡張した手法について述べる。

前述のように λ_0 は document stream 毎に設定され、その値は平常状態における単位時間あたりの document 数と定義される。つまり、実際の計算の際にはこの値が平常状態での document 数を示す値となる。そこで、この値を平常状態での document 数が異なる時間区間毎に変化させることで、このような document stream に対応させることが可能となる。

よって電子掲示板に対する書き込み数が時間帯毎に異なるという問題に対処するためには次のように λ_0 を定義する。電子掲示板の場合、一日周期の書き込み数変動が毎日繰り返されていると仮定できる。そこで、それぞれの時間帯での平均書き込み数を平常状態での書き込み数と考えることで λ_0 の値を決定することができる。つまり、ある電子掲示板に対して一日に n 個の書き込みが実際に観測された場合を考える。一日を N 分割したときのそれぞれの時間区間 i での平均書き込み数を K_i、時間区間 i の長さを T_i とすると、各 i に対する λ_0 は次のようになる。

$$\lambda_{0,i} = \frac{n \times (K_i / \sum K_i)}{T_i} \tag{8}$$

本来の λ_0 が $\lambda_0 = K/T$ であるように、分子は書き込み数 n のうちのいくつが区間 i に対する書き込みであるかの期待値を示し、分母は区間 i の時間的な長さを示している。

blog 記事からなる document stream に対する λ_0 も同様にして定義できる。ただし blog 記事の場合は電子掲示板の書き込み数変動のような周期性がなく過去のデータを利用することができないため、平均値を利用することはできない。そのため、ある話題について集められた対象 blog 記事集合は、blog 記事の全体集合と同じ変動をすると仮定して計算を行う。

対象 blog 記事集合が n 個の blog 記事を含むとする。このとき、blog 記事の存在する時間区間を N 個に分割し、それぞれの時間区間 i での blog 記事集合全体に含まれる blog 記事数を K_i、時間区間 i の長さを T_i とすると、各 i に対する λ_0 が次のように計算できる。

$$\lambda_{0,i} = \frac{n \times (K_i / \sum K_i)}{T_i} \tag{9}$$

電子掲示板の場合と同様に、分子は区間 i で対象 blog 記事集合がいくつの blog 記事を含んでいるかの平常状態での期待値を示し、分母は区間 i の時間的な長さを示している。

以上から明らかであるように、電子掲示板のように変動が周期的であり、その平均値を算出可能である場合には時間区間の長さ T_i を極限まで短くしていくことが可能であるのに対し、blog 記事集合のように過去の情報を利用できない場合は単位時間区間にある程度の幅を与える必要がある。この幅は、イベント発生による document 数増減の影響を受けない程度の長さを確保する必要がある一方で、本質的ではない document 数増減の影響を排除するためにはできるだけ短い値である方が良いと考えられる。

4 実験

前節で提案した手法の有用性を確かめるため、blog と電子掲示板への書き込み数を対象として、ある単語

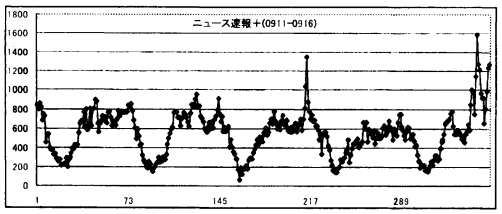

図 3: 「2 ちゃんねる:ニュース速報+」への書き込み数

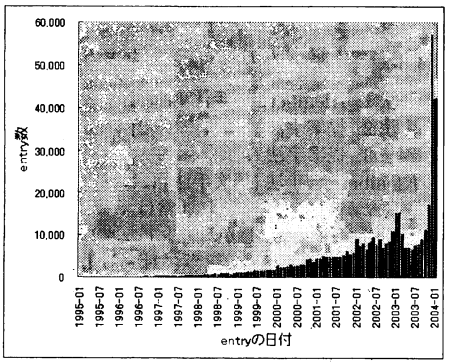

図 4: blog 記事の分布図

が注目状態 (burst) となっているかどうか、ある時期のトピックワードにはどのような語があるかについて実験を行った。

4.1 blog に対する実験

実験対象とする blog は前述のように我々のシステムによって収集された blog 記事である。これは 1995 年 1 月 1 日から 2004 年 1 月 22 日の期間から収集され、466,809 の blog 記事からなっている。また、これらの日付に対する分布数は図 4 のようになっている。そのため document 数の変動を補正することができる提案手法が有効であることが期待できる。なお、提案手法における時間区間の長さ T_i は 1ヶ月として均等に分割した。

まず、ある単語 w が与えられたときにその単語が blog 記事集合中で注目されたことがあったか、また、いつ注目されたかを調べる場合について考える。これは burst 検出手法を用いて以下のように計算することができる。

1. 単語 w を含む blog 記事を集める
2. それらの記事を時系列順に整列する
3. 記事集合を document stream として考え、手法を適用する
4. 各記事が burst 状態であるか、平常状態であるかが結果として得られるので、連続して burst 状態になっている複数の blog 記事からなる期間を、w に関連する話題が注目されていた期間として出力する

また出力された各期間の burst 度スコアとして、その期間に含まれる全記事のスコアの総和を同時に出力する。

表 1: Kleinberg の手法での出力結果:「オリンピック」

burst していた期間	記事数	burst 度スコア
1998-02-06～1998-02-25	34	47.33
2000-08-22～2000-10-14	171	121.4
2002-02-06～2002-03-03	137	84.90
2002-06-03～2002-06-26	30	36.80
2003-02-07～2003-03-06	30	39.61
2003-11-01～2004-01-21	114	138.0

表 2: 提案手法での出力結果:「オリンピック」

burst していた期間	記事数	burst 度スコア
1998-02-06～1998-02-25	34	70.67
2000-08-22～2000-10-14	171	126.1
2002-02-06～2002-03-02	135	67.45

Kleinberg の手法を用いた場合と提案手法を用いた場合、それぞれに対する出力結果として、表 1、表 2 に単語「オリンピック」に対する結果を、表 3、表 4 に単語「クリスマス」に対する結果を挙げる。

次に、blog 記事集合中にどのようなイベントが含まれているかを知るために、月ごとのトピックワードリストの作成を行う。これは、それぞれの月においてどのようなキーワードが注目されていたかを示すランキングであり、その月には、このランキング中のキーワードに関連するイベントが起こっていたと考えることができる。

前述のように、各単語に対する burst を計算することができるので、それを利用して以下の手順でトピックワードの抽出を行う。ただしその際、単語を 1 つだけ表示するのでは、その単語がどのようなイベントを示しているのかが推定できない場合も多い。そこで、その単語と前後に接続する名詞、未知語からなる n-gram を同時に抽出することとする。

1. 単語 w を含む blog 記事を集め、w の burst を計算する
2. ある期間で w が burst 状態になっている場合、w とその期間、burst 度を記憶する
3. 単語 w が burst 状態になっている期間の blog 記事について、w を含む n-gram を全て求める
4. 得られた n-gram の中から w と名詞、未知語の連続部分のみを取り出し、記憶する
5. 全ての単語について繰り返す
6. 月ごとに単語を burst 度の高い順に並べ、n-gram、burst 度と共に出力する

なお、n-gram の計算には工藤らの提案する n-gram に対して拡張された PrefixSpan アルゴリズム [7] を用いた。

出力例を表 5、表 6 に示す。これらの表はそれぞれ Kleinberg の手法を用いた場合と提案手法を用いた場合の、2003 年 11 月のトピックワードをスコアの上位 10 件まで表示したものである。また、n-gram は最大 5 件までを表示し、n-gram 欄に表記されている括弧内の数字はその n-gram の出現回数を示している。

4.2 電子掲示板に対する実験

blog の場合と同様に、電子掲示板に対する実験を行った。ここで対象としたのは電子掲示板群「2 ちゃんね

表 3: Kleinberg の手法での出力結果:「クリスマス」

burst していた期間	記事数	burst 度スコア
1998-12-23～1998-12-26	34	9.724
1999-12-19～1999-12-27	78	19.32
2000-12-10～2000-12-28	176	39.84
2001-11-25～2002-01-02	471	86.94
2002-11-13～2003-01-08	833	137.3
2003-11-05～2004-01-17	3193	201.3

表 4: 提案手法での出力結果:「クリスマス」

burst していた期間	記事数	burst 度スコア
1997-12-18～1997-12-30	24	23.73
1998-12-11～1998-12-27	53	38.28
1999-12-10～1999-12-31	107	47.28
2000-12-10～2000-12-28	176	41.27
2001-12-01～2001-12-29	428	59.93
2002-11-30～2002-12-27	670	58.07
2003-12-09～2003-12-29	2621	49.30

る」[2]中の電子掲示板の一つである「ニュース速報+」に対して、2003 年 12 月 3 日から 2004 年 1 月 3 日の間に書き込まれた 414806 個の書き込みである。この電子掲示板に対する書き込み数は通常、図 3 のように一日周期での増減を繰り返している。そこで提案手法では 2003 年 9 月の一ヶ月間、20 分間隔で書き込み数を調査し、その平均値を平常状態での書き込み数とし、手法中で用いている。

トピックワードリストを blog の時と同様にして計算した結果の上位 10 件を表 7(Kleinberg の手法)、表 8(提案手法) に示す。

5 評価と考察

まず単語「オリンピック」対して 2 手法を適用した結果である表 1、表 2 と、単語「クリスマス」に対する結果の表 3、表 4 について考える。この結果を見ると、Kleinberg の手法の場合では記事数の多い最近の blog に対しては全ての部分で burst となってしまうことがわかる。例えば表 1 の一番最後の burst の場合、オリンピックに関連する目立った話題は存在しないと考えられるにもかかわらず burst となってしまっている。これは記事数が増加したため、単語の出現割合が変化しない場合でも、出現頻度は増加してしまい、その影響を受けているためだと考えられる。一方、表 2 の結果得られた 3 つの burst 期間はそれぞれ長野オリンピック、シドニーオリンピック、ソルトレークオリンピックの時期であることから正しく得られていることがわかる。また、表 3 と表 4 を見比べると、表 3 では burst 度のばらつきがかなり大きくなっている。これも Kleinberg の手法が blog 記事数の増加による影響を受けているためだと考えられる。よって、これらは提案手法が Kleinberg の手法よりもうまく働く典型例であると考えることができる。

次に、blog と電子掲示板のそれぞれから抽出したトピックワードリストを人手で評価することで、手法の有効性を確かめるための評価実験を行った。

blog に関しては、1998 年 2 月 (長野オリンピック)、2000 年 9 月 (シドニーオリンピック)、2002 年 6 月 (サッカーワールドカップ)、2003 年 11 月 (衆議院総選挙)、2003 年 12 月 (計算された中で最新) の 5ヶ月間を評価対象とした。そしてこの 5ヶ月について、以下の 4 つの手法を用いてトピックワードリストを求めた。

- 手法 1: Kleinberg の手法 ($s = 4, \gamma = 2$)
- 手法 2: 提案手法 ($s = 4, \gamma = 2$)
- 手法 3: 提案手法 ($s = 2, \gamma = 1$)
  Kleinberg の手法 (提案手法) には二つのパラメータ s,γ があり、それぞれ通常は $s = 2$,$\gamma = 1$ でよいとされている。しかし、トピックワードリストを作る際には各 blog 記事に対する burst 度を全て足しあわせているため、burst 度が低い単語でも大量に出現している場合にはトピックワードリストの上位に出現してしまう。そのため、s,γ の値を変更し、burst 度が低いものはそもそも burst にならない (burst 度が 0 になる) ようにした方がよい結果が得られるということが予備実験からわかった。予備実験の結果では $s = 4$,$\gamma = 2$ で良い結果を得られたため、手法 1,2 はその値を用い、手法 3 として $s = 2$,$\gamma = 1$ での実験を行う。
- 手法 4: tfidf 法
  その月の各 blog 記事に対し、対象単語に関する tfidf 値を計算し、その総和を対象単語のその月におけるスコアとした。

これらの手法によって計算されたトピックワードから、それぞれ上位 10 件を抜き出して混ぜ合わせることによって作られた各月の単語リストを、1ヶ月につき 2 人の被験者に別々に、各単語がその月のトピックワードとしてふさわしいかどうか評価してもらった。

その結果、それぞれの手法に対する評価は表 9 のようになった。表中のそれぞれの値はその手法がその月のトピックワードとして出力した 10 単語の中でいくつの単語がふさわしいとされたかを示している。なお、括弧外の値は、評価を行った二人のうちのどちらかがふさわしいと判断した単語を正解とした場合の値であり、括弧内の値は二人がともにふさわしいと判断した単語のみを正解とした場合の値となっている。ただし、1998 年 2 月に対して手法 1 ではトピックワードが 2 単語しか得られなかったため、満点は 2 点となっている。この結果からは、どの月に対しても手法 2 が最良の結果を与えていることがわかる。

また、各手法で得られるトピックワードリストの数は表 10 のようになる。なお、手法 4、手法 5 は全ての単語について値を求めることができるため、最大値は全単語になる。この表を見る限り、Kleinberg の手法では、記事数の少ない 1998 年ではトピックワードとして得られる単語数が少なく、逆に記事数の多くなる 2003

表 9: blog:評価結果

	98-02	00-09	02-06	03-11	03-12	平均値
手法 1	1(0)*	7(6)	10(10)	4(2)	2(0)	4.8(3.6)
手法 2	3(2)	7(6)	10(10)	9(9)	7(7)	7.2(6.8)
手法 3	1(1)	5(4)	10(10)	4(2)	2(1)	4.4(3.8)
手法 4	0(0)	1(1)	0(0)	3(1)	1(0)	1.0(0.4)

[2] http://www.2ch.net/

表 5: Kleinberg の手法によるトピックワードリスト

コメント	2003-11-01～2003-11-30	PM コメント (12), 温度風向透明度海状況コメント (11), AM コメント (10), コメントアウト (6), コメント欄 (6)	2752	12196.6
トラック	2003-11-01～2003-11-30	トラックバック (1854), トラックバック 11(12), トラックバック今日 (10), トラックバッド (6), トラックバック日本 (4)	1946	10947.1
バック	2003-11-01～2003-11-30	トラックバック (1855), ライトバック (63), トラックバック 11(12), バックライト (12), バック 11(12)	2098	10900.0
投稿	2003-11-01～2003-11-30	投稿者 (1104), 投稿日 (17), 笑投稿者 (8), shtml 投稿者 (6), 投稿者 myla(5)	1347	7247.4
投票	2003-11-01～2003-11-30	不在者投票 (23), 投票率 (20), 人気投票 (18), 投票日 (9), 投票所 (5)	444	3269.8
選挙	2003-11-01～2003-11-30	選挙区 (21), 衆議院選挙 (19), 日選挙 (14), 議員選挙 (13), 衆議院議員選挙 (12)	324	3026.3
レビュアー	2003-11-01～2003-11-30	30 レビュアー (21), 06 レビュアー (15), 29 レビュアー (13), 24 レビュアー (12), 27 レビュアー (12)	283	2997.7
参考	2003-11-01～2003-11-30	参考書 (14), 参考リンク (13), html.参考 (6), 参考価格 (5), 参考資料 (4)	602	2183.2
レビュー	2003-11-01～2003-11-30	カスタマーレビュー (42), レビュー記事 (8)	379	2029.6
プロフィール	2003-11-02～2003-11-30	プロフィール掲載 (4), プロフィール更新 (3)	216	1541.9

表 6: 提案手法によるトピックワードリスト

単語	burst していた期間	単語周辺の n-gram	記事数	burst 度スコア
選挙	2003-11-07～2003-11-12	日選挙 (14), 選挙区 (14), 衆議院選挙 (13), 衆議院議員選挙 (10), 小選挙区 (6)	187	1361.3
レビュアー	2003-11-06～2003-11-30	30 レビュアー (21), 06 レビュアー (14), 29 レビュアー (13), 13 レビュアー (12), 24 レビュアー (12)	260	954.56
投票	2003-11-07～2003-11-11	投票率 (16), 不在者投票 (13), 投票日 (9), 人気投票 (7), 日投票 (4)	162	794.46
マニフェスト	2003-11-01～2003-11-13	マニフェスト選挙 (5)	49	423.38
小久保	2003-11-03～2003-11-16	小久保裕紀内野手 (7), ダイエー小久保 (5), 日小久保 (4)	39	402.98
レボリューションズ	2003-11-03～2003-11-27	マトリックスレボリューションズ (4)	51	402.92
政党	2003-11-07～2003-11-14	政党化 (3), 政党制 (3)	49	387.60
比例	2003-11-09～2003-11-11	比例区 (11), 比例代表 (10)	39	364.91
衆院	2003-11-01～2003-11-10	衆院選 (33), 衆院選挙 (4)	44	301.00
落選	2003-11-09～2003-11-14		38	298.88

表 7: Kleinberg の手法による「2 ちゃんねる:ニュース速報+」のトピックワードリスト

単語	n-gram	burst 期間	書き込み数	burst 度スコア
お		01/01 00:00～01/01 00:20	434	3823
め	め (33), めこ (24), めこと (19), めめ (5), めー (5)	12/31 23:59～01/01 00:24	376	3054
団塊	団塊 (71), 団塊ジュニア (13), 団塊世代 (10), 団塊 Jr(5), 団塊 Jr.(4)	12/29 17:02～12/29 18:56	261	1679
丼	丼 (23), 牛丼 (12), 豚丼 (9), 焼き鳥丼 (8), 鮭丼 (8)	12/30 18:34～12/30 20:01	214	1469
今年	今年 (85), ー今年 (5), ら今年 (3), 丿) ノボッキアゲ!!今年 (3), め今年 (3)	01/01 00:00～01/01 00:21	175	1371
世代	世代 (19), 教育世代 (7), ゆとり教育世代 (7), 団塊世代 (7), この世代 (5)	12/29 17:02～12/29 18:13	202	1185
IQ	IQ(35), IQ テスト (6), IQ 高く (6), IQ 高い (6), 高い IQ(3)	12/23 10:55～12/23 14:01	180	1108
国連	国連 (20), 国連軍 (10), 国連決議 (7), 国連待機軍 (6), 国連憲章 (3)	12/30 15:46～12/30 16:55	171	1084
愛国心	愛国心 (22), 良い愛国心 (3)	12/24 12:52～12/24 15:01	153	931
岡田	岡田 (36), 岡田幹事長 (6), この岡田 (4), ジャスコ岡田 (4), 岡田克也幹事長 (3)	12/23 21:42～12/23 23:13	132	795

表 8: 提案手法による「2 ちゃんねる:ニュース速報+」のトピックワードリスト

単語	n-gram	burst 期間	書き込み数	burst 度スコア
西村	西村 (48), 西村議員 (19), 西村真悟 (17), 西村氏 (9), 西村先生 (6)	12/19 15:13～12/19 20:11	453	3072
め	め (33), めこ (24), めこと (19), めめ (5), めー (5)	12/31 23:59～01/01 00:24	376	3045
お	お	01/01 00:00～01/01 00:11	329	3044
国連	国連 (46), 国連待機軍 (10), 国連待機部隊 (9), 国連軍 (8), 国連中心主義 (7)	12/30 13:16～12/30 17:04	399	3034
丼	丼 (31), 牛丼 (23), 豚丼 (14), 親子丼 (11), 牛丼中止 (11)	12/30 13:01～12/30 17:38	347	2353
団塊	団塊 (83), 団塊ジュニア (14), 団塊世代 (12), 団塊 Jr(5), 団塊 Jr.(4)	12/29 17:02～12/29 19:19	292	2169
愛国心	愛国心 (43), 愛国心教育 (5), gt; 愛国心 (3), 良い愛国心 (3), ; 愛国心 (3)	12/24 12:23～12/24 17:30	297	1993
世代	世代 (21), 教育世代 (8), ゆとり教育世代 (8), 団塊世代 (7), ジュニア世代 (5)	12/29 17:01～12/29 18:31	243	1895
450 円		12/30 18:09～12/30 20:01	244	1697
宅間	宅間 (39), 宅間さん (7), 宅間守さん (3), 宅間死刑囚 (3)	12/27 12:26～12/27 17:19	218	1396

表 10: blog:得られるトピックワード数

	98-02	00-09	02-06	03-11	03-12	平均値
手法 1	2	25	260	7912	14881	4616
手法 2	23	27	80	36	245	82
手法 3	756	1745	3595	4823	12999	4783

表 11: 「2 ちゃんねる」:評価結果

手法 1	18(17)
手法 2	19(19)
手法 3	1(1)

年では得られる単語数はかなり多くなる傾向にある。それに対し、提案手法では得られる単語数に変化が少ない。つまり、提案手法の方が安定してトピックワードを抽出できていることがわかる。

また、電子掲示板「2 ちゃんねる:ニュース速報+」に対しては表 11 の結果を得た。ここで用いた手法は以下の 3 種類である。

- 手法 1: Kleinberg の手法
- 手法 2: 提案手法
- 手法 3: tfidf 法
  全ての書き込みに対し、対象単語に関する tfidf 値を計算した総和

これらの手法からそれぞれ上位 20 件を抜き出し、二人の人間が評価した結果、表 11 の結果が得られた。しかし、この結果では手法 1,2 の間に精度の面で差は見られなかった。

また手法 1、手法 2 で得られるトピックワード数を時間帯毎に表すと表 12 となる。この表のように、手法 1 では書き込み数の多い夕方から深夜にかけての時間帯からほとんどのトピックワードが選ばれているのに対し、提案手法では書き込み数の少ない昼間からもトピックワードが選ばれている。

6 おわりに

本稿では blog 記事や電子掲示板への書き込みを document stream と考え、Kleinberg の提案する burst 検出手法を拡張した手法を適用することで burst を発見、document stream 中のイベントを表す特徴語の抽出を行った。

提案手法によって得られた結果は、Kleinberg の手法や tfidf 法に比べて良好であると考えられるが、特に電子掲示板においては用意したデータ量が少ないという点に問題があった。また blog に関しても、システムが収集した blog に重複が多く含まれている可能性があり、実際、検出されたトピックワードの中にも一部サイトの影響を強く受けていると考えられるものもあった。そこで今後、これらの問題を解決した状態で改めて評価を行うことが必要であると考えている。

本研究ではトピックワードを含む n-gram を計算し、トピックワードの表すイベントを理解するための助けとして用いた。しかし、提示された n-gram の全てが注目されたイベントに関連しているわけではない。そこで、今後はこのような形で提示するだけではなく、トピックワードとその n-gram のどちらがよりふさわしい

表 12: 「2 ちゃんねる」:得られるトピックワード数

	0~6 時	6~12 時	12~18 時	18~24 時	合計
手法 1	13	4	26	29	72
手法 2	8	3	94	27	132

のかなどを判断できるような手法についても研究することを考えている。

また、本稿ではトピックワード抽出についてのみ述べたが、提案手法はハイパーリンクに対しても適用可能である [8]。そこで、blog 間のリンク関係における burst の発見を行うことで、注目されている話題を別な側面から発見することも可能ではないかと考えている。

謝辞

本研究で利用したデータの一部は、独立行政法人情報処理推進機構 (IPA)「未踏ソフトウェア創造事業」喜連川 優 PM による「blog ページの自動収集と監視に基づくテキストマイニング」に基づくものである。また、本研究は文部科学省科学研究費 (21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」) の補助のもとに行われた。

参考文献

[1] 南野朋之, 鈴木泰裕, 藤木稔明, 奥村学. blog の自動収集と監視. 情報処理学会研究報告, 2003-NL-160, 2004.

[2] Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proc. the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[3] 松村真宏, 大澤幸生, 石塚満. テキストによるコミュニケーションにおける影響の普及モデル. 人工知能学会論文誌 第 17 巻 3 号, pp. 259–267, 2002.

[4] 白倉悟子, 梅村恭司, 小川貴英. 新聞記事における事件特定のための単語群の抽出. 情報処理学会研究報告, 1996-NL-113, 1996.

[5] 仲村大也, 梅村恭司. Katz's k mixture による固有表現の異常出現の検出. 情報処理学会研究報告, 2001-NL-141, 2001.

[6] 長尾真 (編). 岩波講座ソフトウェア科学 15 自然言語処理, pp. 568–576. 岩波書店, 1996.

[7] 工藤拓, 山本薫, 坪井祐太, 松本裕治. 言語情報を利用したテキストマイニング. 情報処理学会研究報告, 2002-NL-148, 2002.

[8] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. of the 12th International World Wide Web Conference*, pp. 568–576, 2003.