情報処理学会論文誌　データベース　Vol.6　No.4　69–84 (Sep. 2013)

情報の注目度とその重要性に基づくトピックの評価指標に関する研究

田中 成典[1]　中村 健二[2]　山本 雄平[3]　柳田 尚明[3,a)]

受付日 2013年3月20日, 採録日 2013年7月5日

概要：CGM の普及にともない，トピックに対する質問，意見，感想や情報提供といったあらゆる反応がインターネットに投稿されるようになった．投稿内容には多様な価値観に基づく情報が含まれていることから，その投稿から有用なものだけを抽出する手法が求められている．既存手法では，バーストの評価指標に基づく注目度合いや，重要性を評価する指標である情報量によって投稿された情報を評価する手法が提案されている．しかし，バーストは投稿件数に基づいた指標のため，投稿内容の重要性が評価できない．また，情報量は投稿の内容に基づいた指標のため，ユーザの注目度合いが評価できない．そこで，本研究では，これら 2 つの指標を組み合わせ，ユーザの注目度合いと投稿内容の重要性に基づきトピックを評価する新たな指標を提案する．そして，評価実験を行うことで本提案指標の有用性を確認する．

キーワード：情報評価指標，トピック情報量，バースト，CGM，データマイニング

Research Concerning Evaluation Indexes of Topics Based on Important Degree of Focused Information

Shigenori Tanaka[1]　Kenji Nakamura[2]　Yuhei Yamamoto[3]　Naoaki Yanagida[3,a)]

Received: March 20, 2013, Accepted: July 5, 2013

Abstract: With the spread of CGM, all kinds of reactions to a topic such as questions, opinions, impressions, and provision of information started to be posted on the Internet. Since information contained in the contents of those posts is based on diverse sense of values, a method for extracting only the useful from the posts is needed. Existing methods suggest approaches to evaluate the degree of drawing attention based on the evaluation index of 'burst', or the posted information according to the amount of information that is an index to evaluate importance. However, burstiness does not help evaluating the importance of the contents of a post, because burstiness is index based on the number of the post. And amount of information does not help evaluating the degree to which it draws users' attention, because amount of information is index based on the contents of a post. There are problems that burstiness does not help evaluating the importance of the contents of a post, and that the amount of information does not help evaluating diverse reactions of users to the post. This study proposes a new index for evaluating a topic according to the degree to which it draws users' attention and the importance of the contents of a post by combining these two indices. And we demonstrate the effectiveness of the proposed index by the demonstration experiments.

Keywords: evaluation indexes of information, amount of topic information, burst, CGM, data mining

[1] 関西大学総合情報学部
Faculty of Informatics, Kansai University, Takatsuki, Osaka 569–1095, Japan
[2] 大阪経済大学情報社会学部
Faculty of Information Technology and Social Science, Osaka University of Economics, Osaka 533–8533, Japan
[3] 関西大学大学院総合情報学研究科
Graduate School of Informatics, Kansai University, Takatsuki, Osaka 569–1095, Japan
[a)] k086121@gmail.com

1. はじめに

SNS（Social Network Service），ブログや掲示板などの CGM（Consumer Generated Media）が普及し，インターネットに流通する情報が増加 [1] している．これらの情報の中には，様々なトピックに対する質問，意見，感想など，消費者や企業にとって有用な情報が多く含まれている．し

かし，その一方で，既知の情報や文字数の少ないユーザの応答情報（たとえば，相づち）など，有用性の低いものも含まれている．そのため，有用な情報を発見するには，利用者自身が膨大な情報から取捨選択する必要があり，多くの時間と労力を要する．

有用性の高い情報を抽出する手法として，バースト解析手法 [2] や LDA（Latent Dirichlet Allocation）[3] を応用したホットトピックの抽出手法 [4] が提案されている．バースト解析手法 [2] は，バーストの有無を判定することで，注目されているトピックや情報の取捨選択が可能である．実際にバースト解析手法は，ブログ解析 [5]，トピック解析 [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] など幅広い分野で応用されている．また，LDA を応用したホットトピックの抽出手法 [4] は，LDA で推定した潜在的なトピックと文書の生成時間に基づく時間フィルタを組み合わせることで，バースト解析手法のみでは抽出できなかった潜在的かつバースト性を有するトピックの抽出を実現している．しかし，文献 [2], [4] の手法では，解析対象のデータを蓄積し，その中で注目度の高いトピックの有無を判定するため，CGM のように新たな情報がリアルタイムに投稿され続けるメディアを対象とした場合の解析は困難である．そのため，リアルタイムに情報が増加するデータストリームの解析に対応した手法としてリアルタイムバースト解析手法 [12], [13], [14], [15], [16] が提案されている．

リアルタイムバースト解析手法 [12], [13], [14], [15], [16] は，バースト解析手法 [2] と異なり，イベントが発生するたびにバーストの有無を判定する．そのため，CGM のようなつねに最新の情報が発生する場合でもバースト解析が可能である．しかし，これらの手法は，バーストの有無をリアルタイムに判定するのみであり，情報そのものの価値を評価していない．そのため，バーストの評価結果に基づき情報を取捨選択した場合には「トピックに対する非難や批判などの誹謗中傷を含む記事」や「コメントや相槌などの短い文章で表現された有用性の低い記事」といった情報そのものに価値がない場合でも評価値が高くなるという問題と，「リーク情報や初期段階のクチコミ情報などのインターネットにあまり流通していない内容を含む記事」といった情報そのものに価値があったとしても，大多数のユーザが発見できていない場合は評価値が低くなるという問題がある．

そこで，本研究では，これら 2 つの問題を解消するために，リアルタイムバースト解析手法の結果に対して，情報の価値を評価する指標を組み合わせることで，情報の重要性を考慮した情報評価指標を提案する．

2. 研究概要

2.1 研究目的

本研究では，インターネットから有用な情報を抽出する際の「情報そのものに価値がない場合でも評価値が高くなるという問題」と「情報そのものに価値があったとしても大多数のユーザが発見できていない場合は評価値が低くなるという問題」を解消することを目的とした新たな情報評価指標を提案する．具体的には，リアルタイムバースト [16] の解析結果と，情報の重要性を評価する指標として一般的である情報量の算出結果とを組み合わせた新たな情報評価指標を提案する．本指標を用いることで，インターネットに流通する情報の有用性の評価が可能となる．

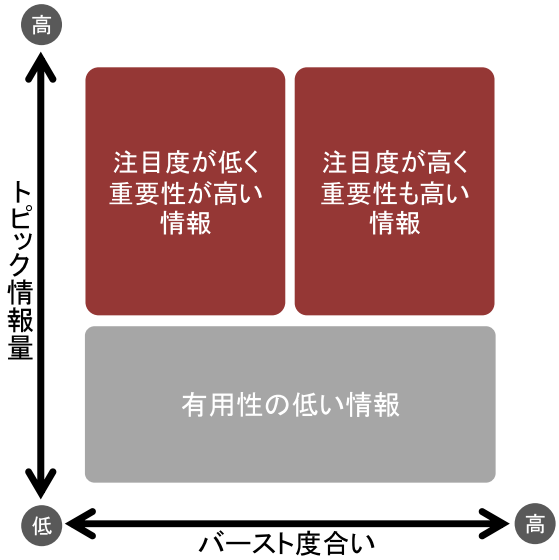

図 1　バーストと情報量の関係

Fig. 1　Relationship between ‘burst’ and ‘amount of information’.

情報量の評価指標として，平均情報量 [17] やカルバック・ライブラ情報量 [18] が，一般的に知られている．本研究では CGM を解析対象としているため，大規模なデータを高速に処理できることが望ましい．そのため，評価値の算出処理がカルバック・ライブラ情報量よりも単純である平均情報量を利用する．

本提案指標では，バーストの解析結果であるバースト度合いと，平均情報量 [17] の考え方をトピックに対応させたトピック情報量とを組み合わせることで，トピックに関する情報の有用性を評価する．これら 2 つの指標に基づき，トピックに関連する情報を評価した場合，**図 1** のように整理できる．図 1 に示すとおり，注目度が高く重要性も高い情報の評価指標を「注目・有用度」，注目度が低く重要性が高い情報の評価指標を「未注目・有用度」と定義し，それぞれの情報を抽出する手法を考案する．

2.2 処理の流れ

本提案指標を利用したシステムの処理の流れを**図 2** に示す．本システムは，インターネットのニュースサイトや掲示板サイトなどを一定間隔ごとに定期監視し，入力されたトピックに関連するキーワード群（たとえば，Facebook，

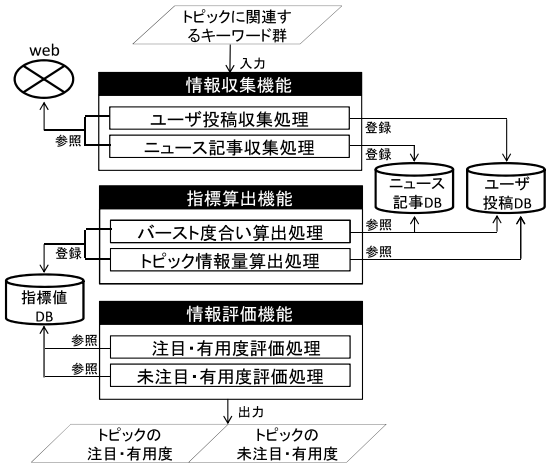

図 2　処理の流れ
Fig. 2　Flow of process.

上場）がタイトルや記事中に含まれる情報を新たに発見すると，その情報を解析して評価値を付与することを想定している．本システムは，情報収集機能，指標算出機能と情報評価機能の 3 つの機能と，ニュース記事 DB，ユーザ投稿 DB と指標値 DB の 3 つの DB とで構成される．本システムの DB は，トピックに関連するキーワード群が入力されるまですべてが空の状態であり，処理が実行されることで逐次情報が格納される．本システムの処理の流れを次に示す．

STEP 1　STEP 1 の処理はキーワード群の入力時とそれ以後の定期監視時に実行される．キーワード群の入力時では，そのときよりも前に投稿されたニュース記事やユーザ投稿のうち，タイトルや記事中にキーワード群が含まれるものを取得する．定期監視時では，キーワード群の入力時での処理と同様に，タイトルや記事中にキーワード群が含まれるニュース記事やユーザ投稿を取得し，それらが前回処理時と比較して新たに投稿されているものかを確認する．ここで，新たな情報が発見されない場合は処理を終了する．

STEP 2　STEP 1 の処理において，ニュース記事やユーザ投稿が取得された場合，STEP 2.1 から STEP 2.3 の処理を実行する．

STEP 2.1　情報収集機能において，サイトごとに事前に登録した正規表現に基づき，投稿日時，タイトルや本文を収集し，それぞれニュース記事 DB とユーザ投稿 DB に登録する．そのため，ニュース記事 DB とユーザ投稿 DB には，キーワード群の入力時とそれ以後の定期監視時に取得したニュース記事やユーザ投稿の投稿日時，タイトルや本文が格納される．

STEP 2.2　指標算出機能において，ニュース記事 DB とユーザ投稿 DB を参照して，STEP 2.1 で新たに登録された情報のトピック情報量とバースト度合いを算出し，指標値 DB に登録する．

STEP 2.3　情報評価機能において，指標値 DB を参照して STEP 2.1 で新たに登録された情報の注目・有用度と未注目・有用度を算出する．

3. 情報の評価アルゴリズム

3.1 概要

本研究では，バースト度合いとトピック情報量とを組み合わせて注目・有用度と未注目・有用度を算出する．バースト度合いは，リアルタイムバースト解析手法 [16] を利用して算出する．また，トピック情報量は，トピックに関連する過去の投稿情報と新たな投稿情報とを利用して算出する．本章では，バースト度合いとトピック情報量の算出方法を説明し，その後，注目・有用度と未注目・有用度の算出方法について述べる．

3.2 バースト度合いの算出方法

バースト度合いとは，バーストの強さを表す指標であり，これを利用することでトピックに関する新たな情報が出現した場合に，そのトピックがどれだけ注目されたかを定量化できると考えられる．本研究で利用するバースト度合いは，既存研究 [16] で提案されたものを用いているため，詳細は文献 [16] を参照されたい．本研究では，トピック t に関して新たに投稿された情報 x のバースト度合いを $Burst(t,x)$，トピック t に投稿された情報のバースト度合いの評価値群を $B(t) = \{Burst(t,1), Burst(t,2), \cdots, Busrt(t,x)\}$ と表す．

3.3 トピック情報量の算出方法

トピック情報量とは，トピックに関連する情報が新たに投稿されたときの情報量の増加分を定量化する指標である．インターネットに流通する情報は，ユーザがブログや掲示板などに投稿する情報と報道機関などの組織が配信する情報が混在していると考えられる．そこで，本研究では，ユーザが投稿する情報を「ユーザ投稿」，報道機関などの組織が配信する情報を「ニュース記事」と定義し，これらの情報が保持する情報量を組み合わせることでトピック情報量を算出する．このとき，ユーザ投稿情報量やニュース記事情報量は，平均情報量 [17] の考え方をトピックに対応させた式 (1) を用いてそれぞれ算出する．N 個の単語で構成された情報 $x = \{w_1, w_2, \cdots, w_k, \cdots, w_N\}$ がトピック t に新たに投稿された場合の情報量 $H(t,x)$ の算出方法を式 (1) に示す．

$$H(t,x) = -\sum_{k=1}^{N} P_{tw_k} \log_2 P_{tw_k} \qquad (1)$$

ここで，P_{tw_k} はトピック t における単語 w_k の出現割合を指しており，P_{tw_k} は平均情報量における確率 P と対応している．式 (1) における P_{tw_k} は，トピックに関連する過

去に投稿された情報と新たに投稿された情報とをそれぞれ構成する単語の異なり語数から算出する．P_{tw_k} の算出方法を式 (2) に示す．

$$P_{tw_k} = \begin{cases} \frac{1}{totalAppear(t)} & (w_k \ is \ new \ word) \\ 1 & (w_k \ is \ not \ new \ word) \end{cases} \quad (2)$$

式 (2) において，$totalAppear(t)$ はトピック t に関連する過去の投稿情報に含まれる単語の異なり語数を指す．式 (2) では，過去の投稿情報の件数が増加するほど，ユーザ投稿情報量やニュース記事情報量が限りなく 0 に近い値となるという問題が発生する．そのため，過去の投稿情報の参照期間を設定するウィンドウサイズ $Wsize$ を導入する．$Wsize$ の期間の投稿情報を用いて $totalAppear(t)$ を算出することで，投稿情報の件数を抑えることができ，この問題の発生を抑制できる．また，単語 w_k が $Wsize$ の期間の投稿情報にも含まれていた場合，単語 w_k は既出単語であるため，その単語の情報量は 0 であると考えられる．そこで，P_{tw_k} の値を 1 にすることで，単語 w_k が保持する情報量 $P_{tw_k} \log_2 P_{tw_k}$ を 0 とする．このようにして算出したユーザ投稿情報量とニュース記事情報量をそれぞれ $H_{User}(t,x)$ と $H_{News}(t,x)$ と表す．

また，インターネットでは，ユーザ投稿とニュース記事の割合は時間やトピックによって流動的に変化する．そのため，トピック情報量におけるユーザ投稿情報量とニュース記事情報量のそれぞれが占める割合も同様に変化すると考えられる．そこで，トピック情報量を占める割合が流動的な変化に対応可能なように，ユーザ投稿情報量とニュース記事情報量を加算した値をトピック情報量と定義する．トピック t に新たに投稿された情報 x のトピック情報量 $H_{Topic}(t,x)$ の算出方法を式 (3) に示す．

$$H_{Topic}(t,x) = H_{User}(t,x) + H_{News}(t,x) \quad (3)$$

本研究では，トピック t に投稿された情報のトピック情報量の評価値群を $H(t) = \{H_{Topic}(t,1), H_{Topic}(t,2), \cdots, H_{Topic}(t,x)\}$ と表す．

3.4　注目・有用度の算出方法

注目・有用度は，注目度が高く重要性も高い情報であるかを評価する指標であり，前述のバースト度合いとトピック情報量とを組み合わせて算出する．しかし，バースト度合いとトピック情報量は尺度が異なる評価指標であるため，これらの値をそのまま利用することは適切でないと考えられる．そのため，それぞれの値を 0 から 1 までの値に正規化する．バースト度合い $Burst(t,x)$ を正規化する方法を式 (4) に示す．

$$Burst'(t,x) = \frac{Burst(t,x) - \min(B(t))}{\max(B(t)) - \min(B(t))} \quad (4)$$

トピック情報量の場合も同様の算出方法で正規化する．このとき正規化したトピック情報量を $H'_{Topic}(t,x)$ と表す．

注目・有用度は，バースト度合いとトピック情報量の両方の値が高いほど，その情報は注目されている有用な情報であるという考えに基づいて算出する．そのため，$Burst'(t,x)$ と $H'_{Topic}(t,x)$ を掛けあわせた値を採用する．トピック t に新たに投稿された情報 x の注目・有用度 $D_{Focused}(t,x)$ の算出方法を式 (5) に示す．

$$D_{Focused}(t,x) = Burst'(t,x) \times H'_{Topic}(t,x) \quad (5)$$

3.5　未注目・有用度の算出方法

未注目・有用度は，注目度が低く重要性が高い情報であるかを評価する指標である．未注目・有用度は前述のバースト度合いとトピック情報量とを組み合わせて算出するため，注目・有用度と同様にこれらを正規化した値である $Burst'(t,x)$ と $H'_{Topic}(t,x)$ を利用する．

未注目・有用度は，バースト度合いが低くトピック情報量が高いほど，その情報は注目されていないが有用な情報であるという考えに基づいて算出するため，$1 - Burst'(t,x)$ と $H'_{Topic}(t,x)$ を掛けあわせた値を採用する．トピック t に新たに投稿された情報 x の未注目・有用度 $D_{Unfocused}(t,x)$ の算出方法を式 (6) に示す．

$$D_{Unfocused}(t,x) = (1 - Burst'(t,x)) \times H'_{Topic}(t,x) \quad (6)$$

3.6　評価値に基づく情報の判定方法

本研究では，注目・有用度または未注目・有用度の評価指標を用いて，情報が有用なものであるかを判定するための閾値 $Stopper$ を設定する．$Stopper$ は，トピックに関する過去の投稿情報の注目・有用度または未注目・有用度の評価値群を利用して算出する．トピック t に関する過去の投稿情報の注目・有用度の評価値群 $DF(t) = \{D_{Focused}(t,1), D_{Focused}(t,2), \cdots, D_{Focused}(t,x)\}$ における，閾値 $Stopper$ の算出方法を式 (7) に示す．

$$Stopper(DF(t)) = \max(DF(t)) \times \alpha \quad (7)$$

式 (7) において，α（$0 \leq \alpha \leq 1$）は閾値を決定するためのパラメータである．未注目・有用度の評価値群の場合も同様の算出方法で閾値を決定する．注目・有用度または未注目・有用度が閾値 $Stopper$ を上回った場合，その情報を有用なものであると判定する．

4.　実験計画と準備

4.1　実験計画

実証実験では，本研究で提案する注目・有用度と未注目・有用度の有用性を証明するために，「実験 1：人工データを用いた既存手法との比較実験」，「実験 2：実データを用いた注目・有用度の評価実験」，「実験 3：実データを用いた未注目・有用度の評価実験」を行う．これらの評価実験は

表 1　実験環境
Table 1　Experiment environment.

OS	Windows7 Professional 32 bit
開発言語	Visual C#
CPU	Intel® Core™ i7-2600 Processor @ 3.40 GHz
メモリ	8 GB

表 1 に示す実験環境で行う．

実験 1 では，平均情報量 [17] を応用したトピック情報量，バースト度合い [16]，LDA を応用したホットトピックの抽出手法 [4] とトピック情報量を組み合わせた指標と，注目・有用度との比較により，情報抽出における注目・有用度の有用性を評価する．なお，実験 1 では，各手法の抽出精度を定量的に比較するため，実データを模して作成した人工データを用いて評価する．本実験において，人工データを用いた理由は，実データの収集対象トピックの選択やトピックに関する正解データ（有用な情報）の選択など，主観的に決定可能な尺度があり，他の手法との比較において恣意性が含まれると考えたためである．

実験 2 では，注目・有用度に基づき抽出した情報を分析することで，リアルタイムバースト解析手法における「情報そのものに価値がない場合でも評価値が高くなるという問題」が解消できるかを検証する．

実験 3 では，未注目・有用度に基づき抽出した情報を分析することで，リアルタイムバースト解析手法の問題点である「情報そのものに価値があったとしても大多数のユーザが発見できていない場合は評価値が低くなるという問題」が解消できるかを検証する．なお，未注目・有用度が判定した情報が有用性の高い情報かどうかはその内容を確認しなければ評価できないため，未注目・有用度では実データを用いた評価実験のみとした．

4.2　実験パラメータの設定

本実験では，リアルタイムバースト解析手法でバースト度合いを算出するときのパラメータ N，β，$Wmin$，$Amin$，$Cmin$ や $Wmax$，トピック情報量算出処理でトピック情報量を算出するときのパラメータ $Wsize$，LDA を応用したホットトピックの抽出手法でバースト度合いを算出するときのパラメータ k，T_1，T_2 や J を用いる．各パラメータについて，次に示すとおり設定した．

4.2.1　リアルタイムバースト解析手法のパラメータ N，β，$Wmin$，$Amin$，$Cmin$，$Wmax$

リアルタイムバースト解析手法では，バースト度合いを算出するために N，β，$Wmin$，$Amin$，$Cmin$，$Wmax$ の 6 つのパラメータを設定する必要がある．本実験では既存研究 [16] にならい，それぞれ $N = 50$，$\beta = 0.4$，$Wmin = 1$，$Amin = 15$，$Cmin = 15$，$Wmax = 1$ とした．

4.2.2　トピック情報量算出処理のパラメータ $Wsize$

トピック情報量算出処理では，トピック情報量を算出するために $Wsize$ のパラメータを設定する必要がある．本実験では事前実験の結果，$Wsize = 30$ とした．

4.2.3　LDA を応用したホットトピックの抽出手法のパラメータ k，T_1，T_2 や J

LDA を応用したホットトピックの抽出手法では，バースト度合いを算出するために k，T_1，T_2 や J の 4 つのパラメータを設定する必要がある．本実験では既存研究 [4] にならい，それぞれ $k = 30$，$T_1 = 7$，$T_2 = 14$，$J = 14$ とした．

4.3　人工データの作成

実験 1 で使用する人工データを作成するために，実データを分析し，本実験で作成する人工データの構成を定義する．そして，その定義に従って人工的にニュース記事とユーザ投稿を作成する．

4.3.1　実データの分析と人工データの構成の定義

作成する人工データの構成を定義するために，事前に収集した実データ（トピック 24 件，ニュース記事 1,757 件，ユーザ投稿 474,569 件）を分析したところ，多くのトピックにおいて，次に示す 2 つの傾向が見られることが分かった．実データの詳細は，4.4 節実データの収集を参照されたい．

- **トピックにはトピックに関連のあるニュース記事やユーザ投稿が出現する．**
- **トピックとは関連のないニュース記事やユーザ投稿（雑談など）が一定の件数で出現する．**

そこで，人工データにおいても同様とするため，本研究では，「トピックに関連のあるニュース記事やユーザ投稿で構成されるトピック」と，「トピックに関連のないニュース記事やユーザ投稿で構成されるノイズ」とを組み合わせた人工データを作成する．なお，ニュース記事やユーザ投稿は，新出単語と既出単語によって構成されていると想定し，人工データを作成するにあたり新出単語として使用する単語群（以下，「新出単語群」と略記）と既出単語として使用する単語群（以下，「既出単語群」と略記）の 2 つを事前に作成する．これら 2 つの単語群の単語は，形態素解析器 MeCab [20] において使用が推奨されている IPA 辞書に収録されるものを使用する．単語群の作成手順を次に示す．

STEP 1　IPA 辞書から品詞が名詞である単語をすべて取得する．

STEP 2　無作為に抽出した 1 万件の単語を新出単語群とする．

STEP 3　STEP 2 の新出単語群を除いた名詞の単語集合から，無作為に抽出した 1 万件の単語を既出単語群とする．

4.3.2　ニュース記事の作成

本実験では，トピックに関連のあるニュース記事と関連のないニュース記事を作成する．なお，トピックに関連のあるニュース記事は，各トピックに少なくとも 1 回以上投稿されるものとする．ニュース記事の作成手順を次に示す．

STEP 1　新出単語群から無作為に抽出した 5,000 件の単語をトピックに出現する新出単語群とし，残りの単語群をトピックに出現しない新出単語群とする．

STEP 2　新出単語出現確率ベクトルを作成する．各単語の出現確率は，IPA 辞書の形態素周辺確率[*1]を採用する．また，ベクトルに用いる単語群は，トピックに関連のあるニュース記事の場合，トピックに出現する新出単語群，関連のないニュース記事の場合，トピックに出現しない新出単語群からそれぞれ取得する．

STEP 3　既出単語出現確率ベクトルを作成する．各単語の出現確率は，新出単語出現確率ベクトルと同様に IPA 辞書の形態素周辺確率を採用する．また，ベクトルに用いる単語群は，トピックへの関連の有無にかかわらず既出単語群から取得する．

STEP 4　ニュース記事の件数を 1 から 15 までの値から無作為に設定する．なお，設定する値の範囲は実データの分析結果により決定した．

STEP 5　ニュース記事の件数分だけ STEP 5.1 から STEP 5.5 の処理を繰り返す．

STEP 5.1　ニュース記事の出現日を無作為に設定する．

STEP 5.2　ニュース記事を構成する単語数 N_{News} を 10 から 2,250 までの値から無作為に設定する．なお，設定する値の範囲は実データの分析結果により決定した．

STEP 5.3　ニュース記事に出現する新出単語の割合 α_{News} を任意に設定する．

STEP 5.4　新出単語出現確率ベクトルから $N_{News} \times \alpha_{News}$ 件の単語を取得し，ニュース記事に出現する新出単語に設定する．

STEP 5.5　既出単語出現確率ベクトルから $N_{News} \times (1 - \alpha_{News})$ 件の単語を取得し，ニュース記事に出現する既出単語に設定する．

4.3.3　ユーザ投稿の作成

トピックに関連のあるユーザ投稿と関連のないユーザ投稿を作成する．トピックに関連のあるユーザ投稿は，トピックに関連のあるニュース記事の投稿日に最も多く投稿され，日数が経過するごとにその件数は減少すると考えられる．そのため，ニュース記事の投稿日におけるユーザ投稿の最大件数を設定し，その日付以降のユーザ投稿の件数を影響力の逓減モデル [19] に基づき，件数が 0 件になるまで順に決定する．トピックに関連のないユーザ投稿は，解析する全期間にわたり無作為に投稿されるように作成する．ユーザ投稿の作成手順を次に示す．

STEP 1　新出単語出現確率ベクトルを作成する．各単語の出現確率は，IPA 辞書の形態素周辺確率を採用する．また，ベクトルに用いる単語群は，トピックに関連のあるユーザ投稿の場合，ニュース記事に出現する新出単語群，関連のないユーザ投稿の場合，新出単語群からそれぞれ取得する．

STEP 2　既出単語出現確率ベクトルを作成する．各単語の出現確率は，新出単語出現確率ベクトルと同様に IPA 辞書の形態素周辺確率を採用する．また，ベクトルに用いる単語群は，トピックに関連のあるユーザ投稿の場合，ニュース記事に出現する既出単語群，関連のないユーザ投稿の場合，既出単語群からそれぞれ作成する．

STEP 3　ユーザ投稿の件数は，次に示す手順に従い決定する．トピックに関連のあるユーザ投稿の場合，ニュース記事の投稿日におけるユーザ投稿の件数を 0 から任意に設定した値 $C_{MaxUserT}$ までの値から無作為に設定し，以降の日付におけるユーザ投稿の件数を影響力の逓減モデル [19] に基づき決定する．なお，影響力の逓減モデルにおける半減期 β は任意に設定する．トピックに関連のないユーザ投稿の場合，0 から任意に設定した値 $C_{MaxUserF}$ までの値から無作為に決定する．

STEP 4　ユーザ投稿の件数に達するまで，STEP 4.1 から STEP 4.4 の処理を繰り返し実施する．

STEP 4.1　ユーザ投稿を構成する単語数 N_{User} を 1 から 661 までの値から無作為に設定する．なお，設定する値の範囲は実データの分析結果により決定した．

STEP 4.2　ユーザ投稿に出現する新出単語の割合 α_{User} を任意に設定する．

STEP 4.3　新出単語出現確率ベクトルから $N_{User} \times \alpha_{User}$ 件の単語を取得し，ユーザ投稿に出現する新出単語を設定する．

STEP 4.4　既出単語出現確率ベクトルから $N_{User} \times (1 - \alpha_{User})$ 件の単語を取得し，ユーザ投稿に出現する既出単語を設定する．

4.4　実データの収集

実験 2 と実験 3 で使用する実データを効率的に収集するために，ニュース記事とユーザ投稿の収集元ドメインを選定する．そして，そのドメインから実データとするニュース記事とユーザ投稿を収集する．

4.4.1　ニュース記事の収集元ドメインの選定

ニュース記事の収集元ドメインを選定する．トピックに

[*1] 形態素周辺確率とは，単語の出現しやすさや他の単語とのつながりやすさを組み合わせることで，単語の形態素になりやすさを確率で表したものである．本研究では，IPA 辞書（https://mecab.googlecode.com/files/mecab-ipadic-2.7.0-20070801.tar.gz から入手）に掲載されているコスト値がその単語の出現しやすさであることから，この値を形態素周辺確率として採用する．

関連のあるニュース記事を効率的に収集するため，「多様なトピックのニュース記事を配信していること」と「ポータルサイトを通してニュース記事を提供していること」の2つの条件に基づき選定したところ，朝日新聞，産経新聞，時事通信，日本経済新聞，毎日新聞と読売新聞の6社が候補として抽出された．

これら報道機関のニュース記事を確認したところ，社説の違いは存在するものの，配信されるニュース記事の多くが重複していることが分かった．そのため，これらの報道機関のうち，いくつかの機関のニュース記事を組み合わせることで，その他の報道機関が配信するニュース記事の内容を網羅できると考えられる．そこで，次に示す選定手順に従い，ニュース記事の収集元とするドメインを決定する．

STEP 1　分析対象のトピック50件を無作為に決定する．

STEP 2　報道機関別ニュース記事網羅性ランキングを作成する．ランキングは，STEP 2.1 から STEP 2.3 の手順で作成する．

STEP 2.1　STEP 1 で決定したトピック50件について，それぞれのトピックごとに，ニュース記事網羅率の高い報道機関を決定する．報道機関ごとのニュース記事網羅率は，STEP 2.1.1 から STEP 2.1.3 の手順で算出する．

STEP 2.1.1　トピック t に関連するニュース記事を報道機関ごとに収集する．

STEP 2.1.2　各報道機関が配信するトピック t に関連するニュース記事の網羅率を算出する．網羅率は，「選定候補すべての報道機関が配信したトピック t に関連するニュース記事群を構成する単語の異なり語数」のうち「報道機関 m が配信したトピック t に関連するニュース記事群を構成する単語の異なり語数」が占める割合（以下，「網羅率」と略記）$Cover(t,m)$ とする．$Cover(t,m)$ の算出方法を式 (8) に示す．

$$Cover(t,m) = \frac{totalAppear(t,m)}{\sum_{k=1}^{6} totalAppear(t,k)} \qquad (8)$$

式 (8) において，$totalAppear(t,m)$ は報道機関 m が配信したトピック t に関連するニュース記事群を構成する単語の異なり語数を表す．なお，$Cover(t,m)$ は，報道機関ごとに算出する．

STEP 2.1.3　STEP 2.1.2 で算出した網羅率が最も高い報道機関を記録する．

STEP 2.2　STEP 1 で決定したトピック50件について，STEP 2.1.3 で記録した報道機関を集計する．

STEP 2.3　STEP 2.2 の集計結果に基づき，網羅性の高い報道機関のランキングを作成する．STEP 2.2 で作成した報道機関のランキングを**図 3** に示す．図 3 において，左側のヒストグラムが STEP 2.2 の集計結果，右側の表が網羅性に基づいた報道機関のランキングを示している．

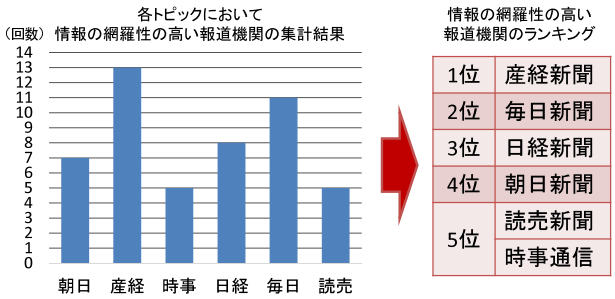

図 3　STEP 2 の結果一覧
Fig. 3　Results by STEP 2.

表 2　統合したニュース記事の累積網羅率
Table 2　Cumulative coverage of integrated news articles.

	最大値	最小値	平均値
産経	0.81	0.30	0.57
産経+毎日	0.92	0.62	0.76
産経+毎日+日経	0.95	0.67	0.85
産経+毎日+日経 朝日	0.98	0.82	0.85
産経+毎日+日経 朝日+時事	1.00	0.89	0.96
産経+毎日+日経 朝日+時事+読売	1.00	1.00	1.00

STEP 3　図 3 のランキングに基づき，上位から順にいくつかの報道機関を組み合わせることで，その他の報道機関が配信するニュース記事を網羅できると考えられる．上位から順に報道機関を組み合わせた値を累積網羅率とし，この累積網羅率が一定値以上となる場合の報道機関の組み合わせを求め，それらを収集元ドメインとして選定する．累積網羅率の算出は，STEP 3.1 から STEP 3.2 で行う．

STEP 3.1　図 3 のランキング上位から順に報道機関を組み合わせたときの累積網羅率を算出する．累積網羅率は，6通りの組合せ（1位の報道機関，1位と2位の報道機関，1位から3位の報道機関，1位から4位の報道機関，1位から5位の報道機関，1位から6位の報道機関）について，それぞれ STEP 3.1.1 から STEP 3.1.3 の処理で算出する．

STEP 3.1.1　組み合わせた報道機関が配信したトピック t に関連するニュース記事を統合する．

STEP 3.1.2　STEP 3.1.1 で統合したニュース記事の累積網羅率を式 (8) を用いて算出する．ただし，式 (8) 中の m は，組み合わせた報道機関群とする．累積網羅率を**表 2** に示す．表 2 は，組み合わせた報道機関の累積網羅率をトピック50件分算出し，累積網羅率が最大，最小となったトピックの値とトピック50件の累積網羅率の平均値を示している．

STEP 3.2　表 2 に基づき，累積網羅率の平均値が 0.80 を超えた際の報道機関の組合せを収集元ドメインとし

て選定する．

まず，図 3 を確認すると，配信するニュース記事の網羅性が高い報道機関は，1 位産経新聞，2 位毎日新聞，3 位日経新聞，4 位朝日新聞，5 位読売新聞と時事通信であることが分かった．次に，表 2 を確認すると，ランキング 1 位から 3 位の報道機関（産経新聞，毎日新聞と日本経済新聞）を組み合わせた場合の累積網羅率が最大値 0.95，最小値 0.67，平均値 0.85 であることが分かった．このことから，これら 3 社を組み合わせることで，その他報道機関である朝日新聞，時事通信と読売新聞のニュース記事の内容をおおむね網羅できることが確認できた．この結果から，本実験では**表 3** に示したドメインをニュース記事の収集元として採用する．

4.4.2　ユーザ投稿の収集元ドメインの選定

ユーザ投稿の収集元ドメインを選定する．トピックに関連のあるユーザ投稿を効率的に収集するため，多様なトピックの情報が活発に投稿される掲示板を選定する．収集元ドメインの選定手順を次に示す．

STEP 1　分析対象のトピック 50 件を無作為に決定する．そして，それらのトピックに関連するキーワード群を検索クエリとして，Google 掲示板検索を行う．

STEP 2　検索結果上位 100 件のドメインを取得する．

STEP 3　ドメインの出現回数を集計し，その上位 20 件を収集元ドメイン（**表 4**）として選定する．

選定したドメインを確認すると，2 ちゃんねる（2ch.net）や FC2 掲示板（bbs.fc2.com）といった大型掲示板サイト，Yahoo!知恵袋（chiebukuro.yahoo.co.jp）やお悩み掲示板（onayamifree.com）といった質問投稿掲示板サイトなど多様なトピックについての情報が活発に投稿されるドメインが取得できていることが分かる．この結果から，各指標の評価実験では，ユーザ投稿の収集元として表 4 に示したドメインを採用する．

表 3　ニュース記事の収集対象とするドメイン
Table 3　Domains for crawling news articles.

報道機関名	収集ドメイン
産経新聞	sankei.jp.msn.com
日本経済新聞	nikkei.com
毎日新聞	mainichi.jp

4.4.3　ニュース記事とユーザ投稿の収集

評価実験で使用する実データでは，解析に用いるトピックに多様性を持たせるため，Yahoo!カテゴリを参考にして 12 のニュースカテゴリを選定する．そして，各カテゴリに対して短期間に集中してニュースが配信されるトピック（以下，「短期トピック」と略記）と定期的にニュースが配信されるトピック（以下，「長期トピック」と略記）とを 1 件ずつ（合計 24 トピック）選定する．各指標の評価実験で用いる短期トピックと長期トピックの一覧を**表 5** に示す．

ニュース記事は，予備実験により選定したドメイン（表 3）からトピックに関連のあるニュース記事を人手で収集し，その見出し，本文と配信日を取得する．ただし，長期トピックは，2009 年 9 月 1 日から 2012 年 8 月 31 日までの 3 年間に投稿された情報に限定して収集する．

各トピックのユーザ投稿は次に示す手順で収集する．

表 4　ユーザ投稿の収集対象とするドメイン
Table 4　Domains for crawling users' posts.

収集ドメイン	出現回数	収集ドメイン	出現回数
2ch.net	272	musyoku.com	50
web2ch.org	260	e-mansion.co.jp	47
groups.google.com	254	machi.to	40
desktop2ch.net	234	onayamifree.com	32
chiebukuro.yahoo.co.jp	172	ezbbs.net	31
jbbs.livedoor.jp	141	bbs.fc2.com	25
qa.itmedia.co.jp	75	2chan.net	25
shizu.0000.jp	73	progoo.com	23
bakusai.com	59	community.teacup.com	11
mikle.jp	50	meiwasuisan.com	11

表 5　実験で利用するトピック一覧
Table 5　Topics using by experiments.

カテゴリ	ID	短期トピック	ID	長期トピック
エンターテイメント	1	実写版るろうに剣心	13	東京ディズニーランド
メディアとニュース	2	東野圭吾ミステリーズ	14	ペプシ，季節限定
趣味とスポーツ	3	第 94 回全国高校野球選手権大会	15	F1，2011
ビジネスと経済	4	Facebook，上場	16	円相場
生活と文化	5	関西大学，レスリング部	17	B-1 グランプリ
芸術と人文	6	劇団四季，CATS	18	芥川賞
コンピュータとインターネット	7	マイクロソフト，Surface	19	著作権法改正
健康と医学	8	福島県，初ガツオ	20	遺伝子組み換え食品
教育	9	いじめ，大津	21	全国学力テスト
政治	10	第 178 回臨時会	22	大阪維新の会
自然科学と技術	11	金井宣茂	23	ノーベル物理学賞
地域情報	12	宇治市，豪雨	24	祇園祭

表 6 「芥川賞」で除去したスレッドと採用したスレッドの例
Table 6 Examples of threads removed and adopted by "Akutagawa award".

	ドメイン	スレッド名
除去	bbs.fc2.com	ブログを作りました
	bbs.fc2.com	冬 到来
	ch-sakura.jp	空母潜水艦
	machi.to	◆下関市 Part24 ◆
	shizu.0000.jp	駿河区石田にある石田神社について
採用	2ch.net	第 147 回 芥川賞・直木賞 候補決定
	2ch.net	田中慎弥氏の受賞会見にネット騒然
	2ch.net	第 145 回芥川賞は該当作なし！
	desktop2ch.net	芥川賞・直木賞の候補作発表
	ezbbs.net	芥川賞選考委員、黒井千次さん退任へ

STEP 1　トピック名と予備実験により選定したドメイン（表 4）を組み合わせて検索クエリを作成する．

STEP 2　作成した検索クエリを用いて Google 掲示板検索を行い，取得した検索結果上位 40 件のスレッドを収集し，そのタイトル，レスの内容と投稿時間を取得する．

STEP 3　トピックに関連のある情報から著しく内容が異なるスレッドを人手で除去する．なお，スレッドの除去作業は，情報関係の有識者 2 人で行い，除去作業に偏りが生じることを避けるため，次に示す手順で実施した．まず，スレッドのタイトルに，トピックに関するキーワード群がないものを除去候補データとして抽出する．次に，1 人目が，目視でトピックに関連する情報であるかどうかを判断し，トピックに関連する情報であった場合は除去候補から除外する．このときの判断基準は，トピックに関する単語の有無とする．最後に，2 人目は，1 人目が作成した除去データを確認し，判断結果が適切であるかどうかを判定する．ここで，判断結果が不適切であると判定されたデータについては，2 人の協議により除去対象とするかを決定する．本作業の一例として，長期トピック「芥川賞」での判断結果を**表 6** に示す．

5. 実験 1：人工データを用いた既存手法との比較実験

5.1 実験内容

本実験では，注目・有用度とその他の指標による情報の抽出精度を比較することで，有用な情報の判定における注目・有用度の有用性を評価する．本実験で用いる手法は，**図 4** に示すとおり，平均情報量 [17] を応用したトピック情報量のみを用いた手法（以下，「情報量手法」と略記），バースト度合い [16] のみを用いた手法（以下，「バースト手法」と略記），LDA を応用したホットトピックの抽出手法 [4] とトピック情報量を組み合わせた手法（以下，「LDA 手法」と略記）と，注目・有用度を用いた手法（以下，「注目・有用度手法」と略記）とする．これら 4 つの手法で算出した評価指標に基づき有用な情報を抽出し，その結果に基づきそれぞれの手法の抽出精度を評価する．

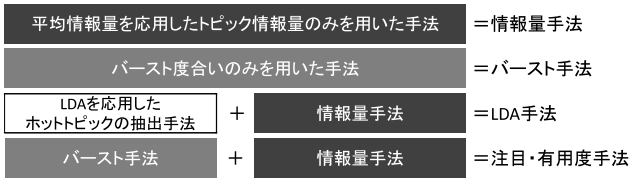

図 4　実験 1 で用いた評価指標
Fig. 4　Evaluation indexes for Experiment 1.

表 7　実験 1 で用いた人工データ作成のパラメータ
Table 7　Parameters for creating artificial data in Experiment 1.

パラメータ	設定値
α_{News}	0.30
α_{User}	0.30
β	3
$C_{MaxUserT}$	100
$C_{MaxUserF}$	30

なお，LDA 手法は，解析対象のデータを蓄積し，その中で注目度の高いトピックの有無を判定する手法であり，そのままでは他の手法と比較できないと考えられる．そのため，本実験では，情報の投稿ごとに判定処理を行うことで対応する．また，情報量手法，バースト手法および LDA 手法では，注目・有用度手法と同様に式 (7) に従い有用な情報を判定する．なお，式 (7) ではパラメータ α を設定する必要があるため，本実験では α と F 値の関係も明らかにすることを目的に，0.00 から 1.00 まで 0.01 刻みで変更してそれぞれ実験を行うことで，α による各手法の抽出精度への影響を確認する．本実験の手順を次に示す．

STEP 1　4.3 節人工データの作成に従い，実験データを作成する．人工データの作成に必要となるパラメータとその設定値を**表 7** に示す．これらのパラメータは，実データを分析した結果をもとに設定した．ただし，LDA の解析時間を短縮するため，トピックに関連のあるユーザ投稿の件数 $C_{MaxUserT}$ を 100 件とし，トピックに関連のないユーザ投稿の件数 $C_{MaxUserF}$ を実データの分析結果より設定した．実データの分析は，2012 年 3 月～2012 年 7 月の各月から無作為に選択した日付のユーザ投稿を用い，それらのデータをトピックへの関連の有無で分類して実施した．その結果，任意のトピックに関連のあるユーザ投稿を 100 とした場合，関連のないユーザ投稿件数は，それぞれ 3 月 = 10，4 月 = 30，5 月 = 30，6 月 = 15，7 月 = 22 となった．本パラメータはユーザ投稿件数の最大値を示すため，$C_{MaxUserF} = 30$ として設定した．これらのパラメー

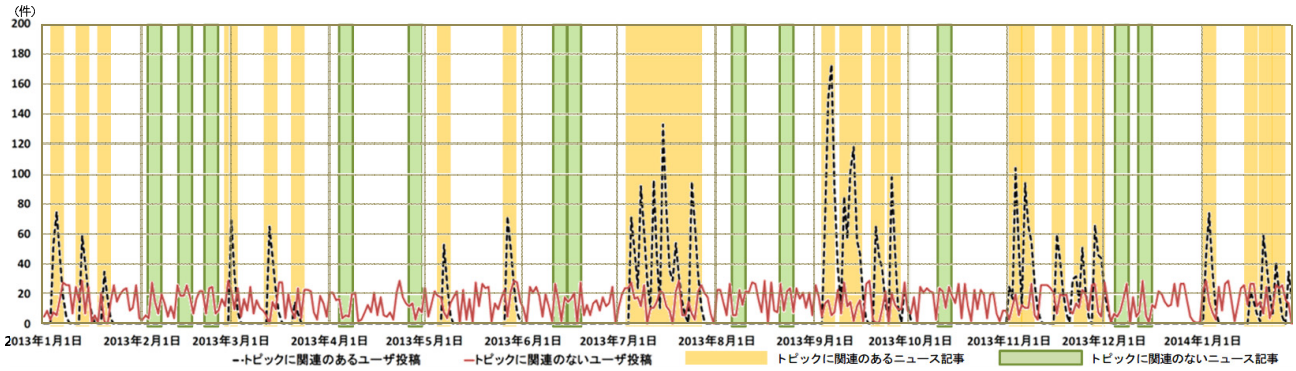

図 5　実験 1 において作成した人工データ
Fig. 5　Artificial data in Experiment 1.

タを用いて作成した人工データを**図 5** に示す．図 5 は，作成したトピックに関連のあるユーザ投稿と関連のないユーザ投稿の件数を日付ごとに表しており，縦軸がユーザ投稿の件数，横軸が日付である．本実験では，定期的にニュースが配信されるトピックを模した人工データを作成するために，トピックが出現していない期間と出現している期間を 1 カ月間隔で交互に繰り返すような特徴を持つ人工データを作成する．なお，作成した人工データでは，トピックに関連のあるニュース記事を正解データとする．

STEP 2　α を 0.01 刻みで変更し，その値が 1.00 になるまで STEP 2.1 から STEP 2.2 の処理を繰り返す．

STEP 2.1　各手法で人工データを解析し，有用な情報と判定したニュース記事の投稿日を抽出する．

STEP 2.2　STEP 1 で作成した人工データにおける正解データの日付と各手法で抽出した日付とを比較し，適合率，再現率と F 値を算出する．

STEP 3　各手法における人工データの平均解析時間を算出する．

5.2　結果と考察

情報量手法，バースト手法，LDA 手法と注目・有用度手法における正解データの抽出精度を**図 6** に示す．図 6 は，α の値を 0.00 から 1.00 まで 0.01 間隔で変更して算出した各手法の F 値を示しており，縦軸が F 値の値，横軸が α の値である．また，各手法における最良の F 値とそのときの α の値を**表 8**，1 回の解析における平均解析時間を**表 9** に示す．これらを確認すると次に示す 2 つの内容が明らかとなった．

- **注目・有用度手法が他の手法よりも高精度に有用な情報を抽出できる**

  表 8 の F 値を確認すると，情報量手法が 0.70（$\alpha = 0.15$），バースト手法が 0.70（$\alpha = 0.39$），LDA 手法が 0.70（$\alpha = 0.13$），注目・有用度手法が 0.77（$\alpha = 0.10$）となり，α の値が異なるものの，注目・有用度手法が他の手法より 0.07 の差で高いことが分かる．そのため，注目・有用度手法とそれぞれの手法との F 値の差が統計的に有意であるかを確認するため，t 検定を実施した．まず，注目・有用度手法と情報量手法とは，等分散であったため，スチューデントの方式による t 検定を実施した結果，$t(200) = 2.11$,

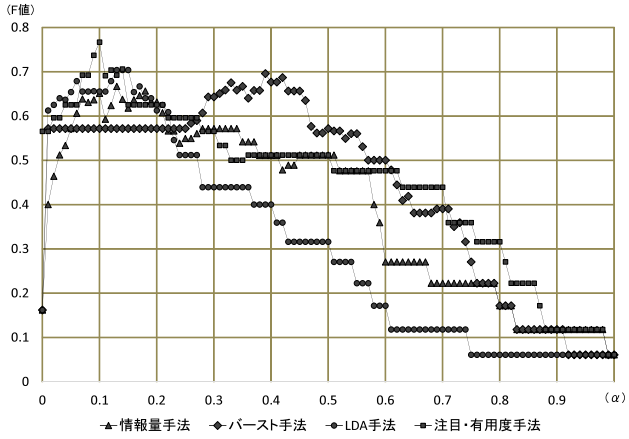

図 6　正解データの抽出精度
Fig. 6　Extraction accuracy of correct data.

表 8　F 値の最大値
Table 8　Maximum value of F-measure.

	α	適合率 (適合数/抽出数)	再現率 (適合数/正解数)	F 値
情報量手法	0.15	0.60(24/40)	0.75(24/32)	0.70
バースト手法	0.39	0.65(24/37)	0.75(24/32)	0.70
LDA 手法	0.13	0.86(19/22)	0.59(19/32)	0.70
注目・有用度手法	0.10	0.82(23/28)	0.72(23/32)	0.77

表 9　平均解析時間
Table 9　Average analysis time.

手法	平均解析時間
情報量手法	00:02.994
バースト手法	00:00.286
LDA 手法	03:13.703
注目・有用度手法	00:03.280

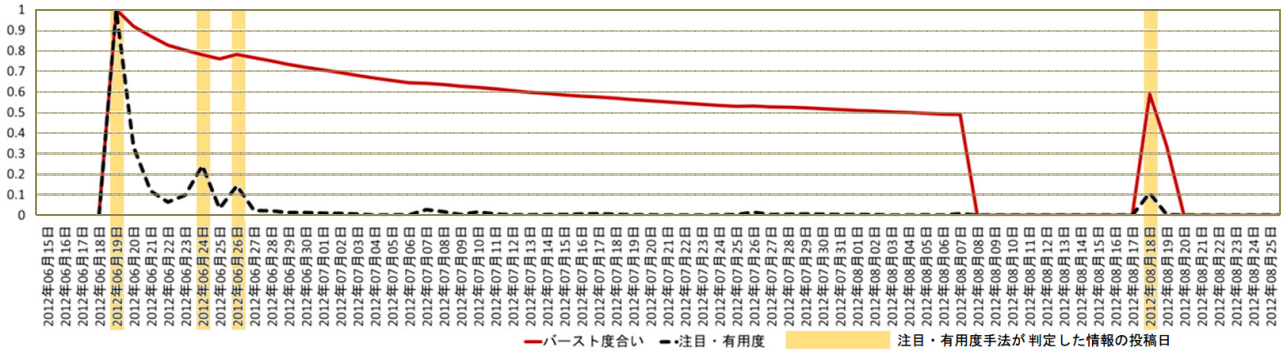

図 7 「マイクロソフト，Surface」の解析結果
Fig. 7 Analysis result about “Microsoft” and “Surface”.

$p < .05$ となった．このことから，注目・有用度手法と情報量手法とは有意差があることが分かった．次に，注目・有用度手法と LDA 手法とは，不等分散であったため，Welch の方式による t 検定を実施した結果，$t(188.831) = 4.77$，$p < .01$ となった．このことから，注目・有用度手法と LDA 手法とは有意差があることが分かった．最後に，注目・有用度手法とバースト手法とは，等分散であったため，スチューデントの方式による t 検定を実施した結果，$t(200) = 0.31, n.s.$ となり，α の値 0.00 から 1.00 までを対象とした場合には有意差がみられないという結果となった．α の範囲を限定して有意差がみられる値を調査したところ，$\alpha = 0.00$ から 0.35 の間では，$t(70) = 2.07$，$p < .05$ となり，2 標本間に有意差がみられる結果となった．このことから，注目・有用度手法の F 値が最良となる $\alpha = 0.10$ を含む $\alpha = 0.00$ から 0.35 の間では，注目・有用度手法とバースト手法とは有意差があることが分かった．

表 8 の適合率に注目すると，LDA 手法が最良で 0.86，注目・有用度手法が 0.82 となり，他の 2 手法よりも約 0.20 ポイント高いことが分かる．一方，再現率に注目すると，情報量手法とバースト手法が最良で 0.75，注目・有用度手法が 0.72 となり，LDA 手法よりも約 0.15 ポイント高いことが分かる．これらのことから，LDA 手法は，正確に情報抽出可能である反面，網羅的に情報を取得できないことが明らかとなった．また，情報量手法とバースト手法は，他の 2 手法と比較して網羅的に情報を取得できる反面，抽出する件数が多い分，正確性に欠ける状況であることが明らかとなった．一方，注目・有用度手法は，適合率と再現率，ともに最良ではないものの，ともに最良の手法とほぼ同等の精度で抽出できており，F 値が最良であることから，他の手法と比較して汎用的に有用な情報を抽出できることが明らかとなった．

- **CGM を対象とした解析には改良が必要であることが分かった**

表 9 を確認すると，情報量手法が 2 秒 994 ミリ秒，バースト手法が 286 ミリ秒，LDA 手法が 3 分 13 秒 703 ミリ秒，注目・有用度手法が 3 秒 280 ミリ秒であることが分かった．これらの処理時間は，各手法の単体のシステムにおける 1 回の処理にかかる計算時間であり，バースト手法が最も高速で，情報量手法，注目・有用度手法が約 3 秒程度かかることが分かった．並列処理や分散処理と組み合わせるためには，それぞれの計算時間を高速化させ，処理に用いるデータを効率的に共有し，各手法での算出処理を細分化して協調させるための改良が必要であることが明らかとなった．

これら 2 つの考察から，CGM を対象として有用性の高い情報を判定するには，高精度かつ高速に情報を抽出できる注目・有用度手法が適していることが明らかとなった．

6. 実験 2：実データを用いた注目・有用度の評価実験

6.1 実験内容

本実験では，実データを対象に注目・有用度手法で抽出した情報を確認することで，その有用性を評価する．なお，本実験では，バースト手法の抽出結果との比較を行うことで，「情報そのものに価値がない場合でも評価値が高くなるという問題」を解消可能であるかを検証する．本実験の手順を次に示す．

STEP 1　トピックに関連のある情報を収集し，実データを収集する．本実験で対象とするトピックは，表 5 に示した 24 トピックとし，解析間隔を 1 日間隔とする．

STEP 2　各手法でトピックを解析し，有用な情報と判定したニュース記事の投稿日を抽出する．

STEP 3　特徴的な反応を示したトピックについて，それぞれが抽出した日付との比較グラフを用いて詳細に分析する．

STEP 4　特徴的な反応を示した日付のユーザ投稿を分類し，トピックに対する有用な情報の有無を確認する．

6.2　結果と考察

実験結果を確認すると次に示す 2 つの内容が明らかとなった．

- **バースト手法における長期間にわたり情報を抽出する現象を抑制できる**

  バースト手法では，長期間にわたり情報を抽出する現象が表 5 の ID5，7，8，9，10，12 のトピックにおいてみられた．ここでは，なかでも特徴的であった「マイクロソフト，Surface（ID7）」の解析結果を用いて考察する．「マイクロソフト，Surface」の解析結果を**図 7**，注目・有用度手法により抽出したデータを**表 10** に示す．図 7 は，バースト度合いと注目・有用度の評価値を日付ごとに表しており，縦軸がそれぞれの評価値，横軸が日付である．図 7 と表 10 を確認すると，バースト手法は 2011 年 6 月 19 日から 2012 年 8 月 7 日まで継続的に情報を抽出していることが分かる．そこで，継続的に抽出した情報を確認すると，雑談や相づちといったノイズ（**表 11**）が多く含まれていることが分かった．注目・有用度手法では特定の日付のみを抽出していることから「情報そのものに価値がない場合でも評価値が高くなるという問題」を解決可能であることが明らかとなった．また，抽出した情報が有用な情報であるかを確認するため，注目・有用度手法で特定した日付のユーザ投稿 100 件を人手で分析し，4 分類に類型化した（**表 12**）．分類結果（表 12）を確認すると，2012 年 6 月 19 日，2012 年 6 月 24 日，2012 年 8 月 18 日のユーザ投稿の約 8 割が，トピックに関連する意見やトピックに関連しないが同様の分野に関する意見であり，有用な情報が抽出できていることが明らかとなった．

  一方，2012 年 6 月 26 日に抽出した情報を確認すると有用性の低い情報を抽出していることが分かる．ユーザ投稿の分類結果（表 12）を確認しても，約 5 割が関係のない話題やノイズであり，あまり有用ではない情報が抽出されている状況である．2012 年 6 月 26 日が抽出された原因を確認すると，図 7 に示されているとおり，この日付は，バースト度合いの値が前日よりも高い値を示した日付であることが確認できる．このことから，これら情報の抽出を抑制できなかった原因は，高い値を示したバースト度合いに注目・有用度が影響を受けたことにより，バーストしている状態を十分に抑制できなかったためであると考えられる．これについては，注目・有用度の反応を判定する際にトピック情報量が一定値以下の場合に反応を抑制することで解消できると考えられる．

- **バースト手法において断続的に情報を抽出する現象を抑制できる**

  バースト手法では，断続的に情報を抽出する現象が表 5 の ID1，13，14，16，17，18，19，23 においてみられた．ここでは，なかでも特徴的であった「B-1 グランプリ（ID17）」の解析結果を用いて考察する．「B-1 グランプリ」の解析結果を**図 8**，注目・有用度手法が抽出したデータを**表 13** に示す．図 8 は，バースト度合いと注目・有用度の値を日付ごとに表しており，縦軸がそれぞれの評価値，横軸が日付である．図 8 と

表 10　「マイクロソフト，Surface」により抽出されたデータ
Table 10　Data extracted by “Microsoft” and “Surface”.

日付	情報（収集元ドメイン）
2012/06/19	Microsoft、タブレット端末「Surface」を発表（nikkei.co.jp）
2012/06/24	Microsoft の Surface、フリーズ連発で顔真っ赤（2ch.net）
2012/06/26	Microsoft の Surface 関連の雑談（2ch.net）
2012/08/18	タブレット Surface なんと 199 ドル（2ch.net）

表 11　抽出したノイズの例
Table 11　Examples of extracted noise.

ドメイン	ノイズ
desktop2ch.net	それじゃあバイバイ
desktop2ch.net	ワロタ w
desktop2ch.net	イギリス発音だとスーフスに聞こえる
musyoku.com	通信費もっと下げろ
musyoku.com	あ～なる程

表 12　「マイクロソフト，Surface」に関するユーザ投稿の分類
Table 12　Classification of users’ posts concerning “Microsoft” and “Surface”.

	6 月 19 日	6 月 24 日	6 月 26 日	8 月 18 日	具体例
トピックに関連する意見	64	51	33	79	・居間に 1 つ置いて、天気予報や写真を表示させておきたい。 ・みんなには悪いけど俺にとってはすごく魅力的だ
トピックには直接関連のない意見	22	29	21	17	・クラムシェル型欲しいね ・マイクロソフトも UNIX ベースにすりゃいいんだ・・・
その他	4	11	29	2	・そういえば 8 で XBOX のソフトが動くという話があったな。 ・その辺均衡状態に持っていけるかが企業の力の見せ所かなぁ
ノイズ	10	9	17	2	・パクリパクリパクリパクリパクリパクリパクリパクリパクリ ・記念カキコしとこう

表 13 を確認すると，バースト手法では 2011 年 9 月 11 日から 13 日，15 日から 18 日，25 日から 26 日，2011 年 10 月 7 日から 8 日，2011 年 11 月 4 日から 6 日の期間に断続的に発信された情報を抽出していることが分かる．

一方，注目・有用度手法では 2011 年 11 月 12 日から 14 日の期間のみの情報を抽出しており，バースト手法で抽出した情報は抽出していないことが分かる．また，2011 年において B-1 グランプリは 11 月 12 日と 13 日の 2 日間開催されており，注目・有用度手法が抽出した情報の日付とほぼ一致していることから「情報そのものに価値がない場合でも評価値が高くなるという問題」を解決可能であることが明らかとなった．

また，抽出した情報が有用な情報であるかを確認するため，「マイクロソフト，Surface」のトピックを解析した際と同様に，注目・有用度手法で特定した日付のユーザ投稿 100 件を人手で分析し，4 分類に類型化した（**表 14**）．分類結果（表 14）を確認すると，ノイズは含まれるものの 2011 年 11 月 12 日，2011 年 11 月 13 日，2011 年 11 月 14 日のユーザ投稿の約 7 割が，トピックに関連する意見やトピックに関連しないが同様の分野に関する意見であり，有用な情報が抽出できていることが明らかとなった．

表 13　「B-1 グランプリ」により抽出されたデータ
Table 13　Data extracted by “Grand prix of B-1”.

日付	情報（収集元ドメイン）
2011/11/12	姫路にズラリ、B—1 グランプリ開幕（nikkei.co.jp）
2011/11/13	B—1 グランプリ、「ひるぜん焼そば」優勝（2ch.net）
2011/11/14	B—1 グランプリ、ホルモンうどんがまさかの敗北（2ch.net）

これら 2 つの考察から，注目・有用度手法でトピックを解析することで，バースト手法で発生する過剰な反応を抑制しつつ，有用な情報を実データから抽出できることが確認できた．このことから「注目度が高く重要性も高い情報」を抽出するための指標として注目・有用度が適していることが明らかとなった．

7. 実験 3：実データを用いた未注目・有用度の評価実験

7.1 実験内容

本実験では，実データを対象に未注目・有用度手法で抽出した情報を確認することで，その有用性を評価する．なお，本実験では，バースト手法の抽出結果との比較を行うことで，「情報そのものに価値があったとしても大多数のユーザが発見できていない場合は評価値が低くなるという問題」を解消可能であるかを検証する．本実験で対象とするトピックは「関西大学，レスリング部」とし，解析間隔

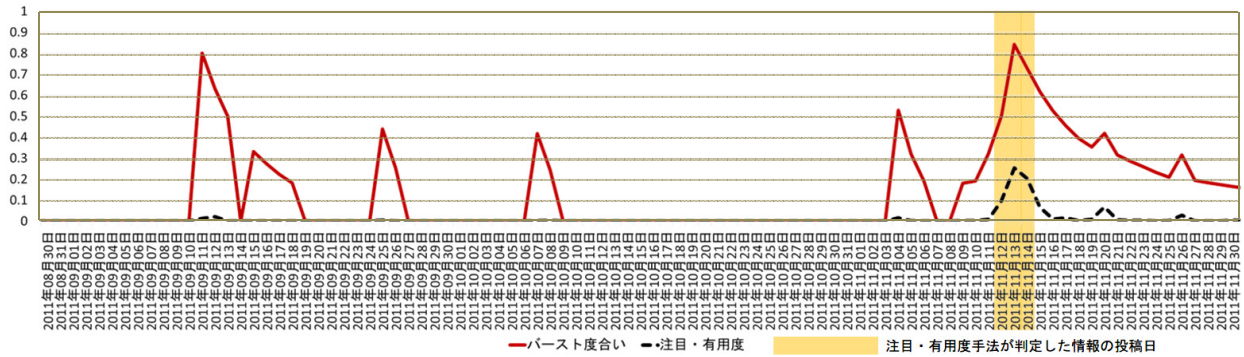

図 8　「B-1 グランプリ」の解析結果
Fig. 8　Analysis result about “Grand prix of B-1”.

表 14　「B-1 グランプリ」に関するユーザ投稿の分類
Table 14　Classification of users’ posts concerning “Grand prix of B-1”.

	11 月 12 日	11 月 13 日	11 月 14 日	具体例
トピックに関連する意見	69	25	31	・現在シロコロ 30 人ほど、横手焼きそば 9 人の列。 ・この勝負方法では製造に時間が掛かるものは勝てない。
トピックには直接関連のない意見	17	43	38	・あんこ嫌いだからきびだんごの方が好き という人も結構いる ・第 2 回の富士宮と第 4 回の横手は地元が優勝しているんだね
その他	3	11	9	・関西から名古屋来たけど何で名古屋って不味いものばっかなの？ ・例えばサッカーのファジアーノ岡山。県北では誰も話題にしてない
ノイズ	17	43	38	・ありがとう！ ・え？

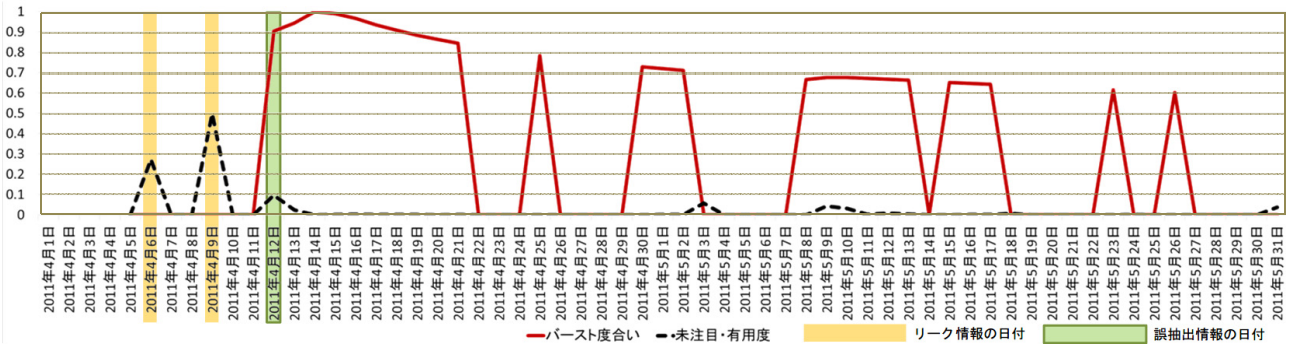

図 9　「関西大学，レスリング部」の解析結果
Fig. 9　Analysis result about “Kansai university” and “Wrestling club”.

表 15　「関西大学，レスリング部」により抽出された情報
Table 15　Data extracted by “Kansai university” and “Wrestling club”.

日付	情報（収集元ドメイン）
2011/04/06	関西大学レスリング部イジメ事件について [質問]（chiebukuro.yahoo.co.jp）
2011/04/09	関西大学レスリング部イジメ事件について [回答]（chiebukuro.yahoo.co.jp）
2011/04/12	部員に加熱トング 関大レスリング部元主将ら捜査（mainichi.jp）

を 1 日間隔とする．「関西大学，レスリング部」のトピックは，2011 年に発覚した不祥事の情報が記者発表前にインターネットにリークされた経緯がある．そのため，未注目・有用度手法で，そのリーク情報が抽出できるかを評価する．

7.2　結果と考察

「関西大学，レスリング部」の解析結果を**図 9** に示す．図 9 は，バースト度合いと未注目・有用度の評価値を日付ごとに表しており，縦軸がそれぞれの評価値，横軸が日付である．この結果を確認すると，次に示す内容が明らかとなった．

- **バースト手法では抽出できない情報を抽出できる**

  バースト手法と未注目・有用度手法により抽出した情報を時系列に沿って確認すると，未注目・有用度手法が 2011 年 4 月 6 日と 2011 年 4 月 9 日に投稿された情報を抽出した後，2011 年 4 月 12 日の情報を両手法ともに抽出していることが分かる．これら抽出した日付に投稿された情報を**表 15** に示す．表 15 を確認すると，未注目・有用度手法が抽出した情報は，2011 年 4 月 12 日の記者発表前にリークされた情報と一致していることが分かる．このことから，バースト手法では抽出できない「注目度が低く重要性が高い情報」を抽出できることが明らかとなった．

  一方，2011 年 4 月 12 日に注目すると，バースト手法だけでなく未注目・有用度手法も情報を抽出していることが分かる．この情報は，記者発表によって一般に公開された情報であることから，この情報は「注目度が高く重要性も高い情報」であると考えられる．そこで，この情報のバースト度合いとトピック情報量の値を確認するとそれぞれ 0.9 と 1.0 であることが確認できた．未注目・有用度がトピック情報量とバースト度合いの差によって算出されることから，それぞれの評価値がともに高い値を示した場合であっても，「注目度が低く重要性が高い情報」として誤抽出する可能性があることが明らかとなった．これについては，未注目・有用度の反応を判定する際にバースト度合いが一定値以上の場合に反応を抑制することで解消できると考えられる．

この考察から，未注目・有用度手法でトピックを解析することでバースト手法の問題点である「情報そのものに価値があったとしても大多数のユーザが発見できていない場合は評価値が低くなるという問題」を解決できることが確認できた．このことから，「注目度が低く重要性が高い情報」を抽出するための指標として未注目・有用度が適していることが明らかとなった．

8.　おわりに

本研究では，情報の重要性を考慮した情報評価指標として，「情報そのものに価値がない場合でも評価値が高くなるという問題」を解消する注目・有用度と，「情報そのものに価値があったとしても大多数のユーザが発見できていない場合は評価値が低くなるという問題」を解消する未注目・有用度とを提案した．

評価実験の結果，有用な情報の抽出に利用可能な他の指標（トピック情報量，バースト度合い，LDA を応用したホットトピックの抽出手法とトピック情報量を組み合わせた指標）よりも注目・有用度の方が，汎用的に有用な情報を抽出できることを立証した．また，注目・有用度を用いることで，バースト度合いを用いた場合の誤抽出を抑制したうえで，注目度合いと重要性が高い情報を抽出できること

を立証した．一方，未注目・有用度を用いることで，バースト度合いでは抽出できなかった注目度が低く重要性が高い情報を抽出できることを立証した．注目・有用度は単語と投稿時間を使用して解析する指標であるため，評価実験で対象とした掲示板だけでなく，ブログ，クチコミサイトや組織のニュースリリースなど，単語と投稿時間を持つ多種多様な情報への適応も期待できる．

今後の課題として，「情報の評価アルゴリズムの改善」，「パラメータ設定の自動化」と「実社会での運用を想定したリアルタイム処理の検討」があげられる．

1つ目の課題は，注目・有用度ではバースト度合いが高い場合，その値に影響を受け，注目・有用度も高い状況となった．これは，バーストしている状態を十分に抑制できなかったために発生しており，有用ではない情報を誤抽出する事例がみられた．また，未注目・有用度はトピック情報量とバースト度合いの差によって算出されることから，それぞれの評価値がともに高い値を示した場合であっても「注目度が低く重要性が高い情報」として誤抽出する事例がみられた．これらの課題に対しては，情報抽出の判定時に，注目・有用度の場合はトピック情報量に対して，未注目・有用度の場合はバースト度合いに対して，それぞれ閾値を設定可能なように拡張することで対応する予定である．

2つ目の課題は，注目・有用度や未注目・有用度による評価値を算出するには8つのパラメータを適切に設定する必要があるという問題である．パラメータ N は，短期トピックと長期トピックで最適な解析期間が異なると考えられるため，解析対象のトピックが短期か長期かを判定し，短期であれば解析期間を短く，長期であれば長くするといった方法が考えられる．パラメータ β は，通常とは異なる評価値を示した情報を抽出することが望ましいと考えられるため，初期段階に学習期間を設け，その期間中の評価値群の平均値といった代表値を採用する方法が考えられる．また，パラメータ α，$Wmin$，$Amin$，$Cmin$，$Wmax$，$Wsize$ はカテゴリごとに最適値が異なるため，投稿パターンや投稿件数といった傾向を分析することでパラメータ設定を自動化できると考えられる．

3つ目の課題は，実社会での運用を想定した場合，大規模なデータ群の解析をリアルタイムに処理できる仕組みが求められるという問題である．この課題に対しては，リアルタイムに対応したデータ構造の考案と評価値の算出処理の細分化と独立化を図り，並列分散処理の考え方を組み込むことで，評価値算出処理全体の高速化を実現し，実データを対象に実験を行うことで対応する予定である．

今後，上記3点の課題への対応と，様々なトピックに対しての解析結果を分析することでこれら指標の改良を目指す．

参考文献

[1] 情報通信政策研究所：我が国の情報通信市場の実態と情報流通量の計量に関する調査研究結果，総務省（オンライン），入手先 ⟨http://www.soumu.go.jp/main_content/000124276.pdf⟩（参照 2013-07-18）．

[2] 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).

[3] Blei, D., Ng, A. and Jordan, M.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022, JMLR (2003).

[4] 水田昌孝，熊野雅仁，小野景子，木村昌弘：文書ストリームからのバースト潜在トピック抽出における t-LDA 法の性能検証，情報処理学会バイオ情報学研究会研究報告，Vol.23, No.10, pp.1–6, 情報処理学会 (2010).

[5] 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).

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

[7] 木村　学，斉藤和巳，上田修功：確率モデルに基づく文書ストリームからのホットトピック抽出の一検討，電子情報通信学会人工知能と知識処理研究会技術研究報告，Vol.106, No.38, pp.51–56, 電子情報通信学会 (2006).

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

[9] He, Q., Chang, K., Lim, E. 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, K.: A Web Personalizing Technique Using Adaptive Data Structures: The Case of Bursts on Web Visits, *Journal of Systems and Software*, Vol.83, pp.2200–2210, Elsevier (2010).

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

[13] Shasha, D. and Zhu, Y.: High Performance Discovery in Time Series: Techniques and Case Studies, pp.151–174, New York University (2004).

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

[15] 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, 日本データベース学会 (2010).

[16] 蝦名亮平，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, pp.86–96, 情報処理学会 (2012).

[17] Shannon, C.: A Mathematical Theory of Communication, *The Bell System Technical Journal*, Vol.27,

pp.379–423, Bell Laboratories (1984).
[18] Kullback, S. and Leibler, A.: On Information and Sufficiency, *Annals of Mathematical Statistics*, Vol.22, No.1, pp.79–86, Institute of Mathematical Statistics (1951).
[19] 石川佳治，北川博之：忘却の概念に基づく文書クラスタリング手法の改良方式について，情報学基礎研究会研究報告，Vol.2003, No.112, pp.1–7, 情報処理学会 (2003).
[20] Kudo, K., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. 2004 Conference on Empirical Methods in Natural Language Processing*, pp.230–237, ACL (2004).

田中 成典　（正会員）

1986 年関西大学工学部土木工学科卒業．1988 年関西大学大学院工学研究科土木工学専攻博士課程前期課程修了．同年，（株）東洋情報システム（現在，TIS）に入社．人工知能に関する研究受託開発業務に従事．1994 年関西大学総合情報学部専任講師として着任，1997 年助教授，2004 年教授，2006 年から学生センター副所長，現在に至る．2002 年 8 月から 1 年間，カナダの UBC にて客員助教授．博士（工学）．専門は知識工学と社会基盤情報学．CAD/CG，GIS/GPS，画像処理および Web ソリューションビジネスに関する研究に従事．2000 年（株）関西総合情報研究所を起業，設立当初から現在まで取締役会長．2006〜2012 年（株）フォーラムエイトの顧問．建設省土木研究所 CAD 製図基準検討委員会委員長，土木学会土木情報システム委員会幹事長，同委員会土木 CAD 小委員会委員長，ISO/TC184/SC4 国内委員等を歴任．現在，国土交通省日本建設情報総合センター社会基盤情報標準化委員会委員，同委員会 CAD/データ連携小委員会委員長，土木学会情報利用技術委員会副委員長．主に，ISO に準拠した CAD 製図基準と CAD データ交換基盤の開発に従事．

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 年関西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 年関西大学大学院総合情報学研究科総合情報学専攻博士課程後期課程修了．同年関西大学ポスト・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，Web マイニング，テキストマイニング等の研究に従事．2002 年から（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．電子情報通信学会，土木学会，日本データベース学会，日本知能情報ファジィ学会各会員．

山本 雄平　（学生会員）

1986 年生．2009 年関西大学総合情報学部総合情報学科卒業．2011 年関西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．現在，関西大学大学院総合情報学研究科総合情報学専攻博士課程後期課程在学中．修士（情報学）．Web マイニング，自然言語処理の研究に従事．2007 年（株）関西総合情報研究所入社．現在に至る．システム設計等の研究開発に従事．

柳田 尚明

1988 年生．2011 年関西大学総合情報学部総合情報学科卒業．2011 年関西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．現在，関西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在学中．学士（情報学）．Web マイニング，自然言語処理の研究に従事．

（担当編集委員　上田 真由美）