情報処理学会論文誌 Vol. 49 No. 7 2480–2492 (July 2008)

GAを用いたWebニュースの時系列情報を考慮したトピック抽出に関する研究

中村 健二[†1] 田中 成典[†2] 古田 均[†2]
吉村 智史[†1] 北野 光一[†1]

近年，インターネットを利用した日常的な情報収集活動において，即時性と信頼性に優れたWebニュースは人々の情報源として広く活用されている．しかし，日々時々刻々と増加する膨大な数のWebニュースから特定のトピックだけを抽出することは困難である．そのため，文書間の類似度を利用してトピックを分類する研究や時系列的な特性に基づきトピックを抽出する研究が活発に行われている．しかし，これら既存研究では，文書中に出現する単語群に依存した分類しかできず，また任意の期間に発生するトピックに適切な単語を関連付けできないという課題がある．そこで，本研究では，時系列的な特性に基づいて抽出したバースト語を用いて，バースト語間の関連を考慮した最新のトピックを抽出する手法を提案する．そして，既研究の従来手法と比較実験を行い，本提案手法の有用性を実証する．

Research for Extracting Topic of Web News with Time Series Information Using GA

Kenji Nakamura,[†1] Shigenori Tanaka,[†2]
Hitoshi Furuta,[†2] Satoshi Yoshimura[†1]
and Kouichi Kitano[†1]

A Web News that is very immediate and reliable is used as a news resource at practical activity of collecting information on Internet in recent years. However, it is difficult to extract specific topics from huge Web News that is increasing momentarily. Therefore, there are some researches which extract the topics depend on a feature of time series information or classified topics using the similarities between documents. However, these past researches have some problems. Firstly, Web News is classified by term frequency in the document set. Secondly, topics that occur over any term can not have links with words that are relevant to a unique topic. Thus, in this paper, we propose a method to extract hot topics based on a relation between burst words that are extracted based on a feature of time series information. Then, we evaluate the effectiveness to compare the proposed method with the existing methods of early researches.

1. はじめに

近年，情報技術の普及と発展にともない，インターネット上に膨大な情報が蓄積され，誰もが日常的な情報収集活動にインターネットを利用[1]している．特に，インターネット上の情報の中でも，新聞社の配信するWebニュースは，現実世界で起きた最新のニュースを瞬時に配信する即時性と，校閲された文章であるために信頼性に優れているという特徴があり，多くの人々の情報源として活用されている．また，Webニュースは時間情報を有しており，時間の経過とともにトピックの発生と消滅を繰り返すという特徴がある．

このような特徴を活かし，任意の期間に発生する複数のトピックを自動抽出し，各トピックに属する文書を正確に整理分類することができれば，日々，時々刻々と配信される膨大なWebニュースから特定のトピックだけを選定できるようになり，必要な情報を検索するためにかかるコストを大幅に削減できる．

ニュースからトピックを抽出する研究は，米国のDARPA（Defense Advanced Research Projects Agency）を中心とするTDT（Topic Detection and Tracking）[2)–5)]のプロジェクトにおいて活発に行われている．TDT[2)]では，時系列情報を持つWebニュースを蓄積したコーパスを対象として，トピック抽出に関する複数の課題を設定している．TDTの課題の1つであるトピックの抽出に関する研究には，学習データを用いて単語の生成確率を事前に推定し，その結果を利用してトピックを抽出する手法と，学習データを用いずに文書集合からトピックを抽出する手法がある．前者の手法には，Naive Bayes法[6)]を基にしたLDA（Latent Dirichlet Allocation）[7)]などがある．LDAは，確率的トピックモデルを用いることで多重トピックを取り扱えるという特徴があり，時系列情報を保持する文書に対応するために拡張した研究[8)]も提案されている．しかし，LDAには，複数のモデルパラメータがあり，そのパラメータ値を決めるために，文書集合から単語の生成確率を生成して学習[9)]さ

†1 関西大学大学院総合情報学研究科
Graduate School of Informatics, Kansai University
†2 関西大学総合情報学部
Faculty of Informatics, Kansai University

せる必要がある．さらに，トピック抽出の精度は学習に利用するデータの量および質に依存するという特性がある．一方，後者の手法は，分類対象が不明瞭であることから事前に学習データを準備することが困難な場合に有効な手法として様々な研究[10)–17)]がなされている．学習データを用いずにトピックを抽出する手法に関する研究では，k-means 法[6)]を拡張してトピックを抽出する研究[10),11)]と，時系列的な特性に基づきトピックを抽出する研究[12)–17)]がある．

本研究では，インターネット上の最新ニュースからのトピック抽出を目的とするため，学習データを用いずにトピックを抽出する手法に着目する．

k-means 法[6)]を拡張してトピックを抽出する研究[10),11)]では，文書に出現する単語の特徴を用いて，文書間の類似度や距離を算出し，その結果により文書集合を複数のトピックに分類する．具体的には，まず，ユーザが分類するトピックの数を決定する．次に，ユーザが指定した数の文書を無作為に選択し，各トピックの中心文書とする．そして，各トピックの中心文書とその他の文書の類似度を利用して，各文書の分類を行う．各トピックの構成文書に変化がなくなるまでこの一連の処理が繰り返される．

時系列的な特性に基づきトピックを抽出する研究[12)–17)]では，時間情報を持つ文書集合から特定の期間に頻出する単語を抽出する．Kleinberg の手法[13)]では，特定の期間に任意の単語を含む文書が頻出することをバースト，そして頻出する単語をバースト語，単語が頻出する期間をバースト期間と定義されている．そして，バースト語を含む任意の文書集合をトピックとして抽出することができる．

しかし，これらの研究には次のような問題点がある．

- 前者の k-means 法を拡張した手法では，文書数の少ないトピックを抽出できない．
- 前者の k-means 法を拡張した手法では，異なるトピックにおいて文書の類似度が高い場合に適切にトピックを抽出できない．
- 後者の時系列的な特性に基づく手法では，文書集合から適切にトピックを抽出できない．

そこで，本研究では，時系列情報に基づいて抽出したバースト語を用いて，バースト語間の関連を考慮した最新のトピックを抽出する手法を新たに提案する．そして，学習データを用いずにトピックを抽出する手法に関する研究で一般的に用いられる k-means 法と Kleinberg の手法と比較実験を行い，本提案手法の有用性を実証する．

2. 研究の概要

2.1 研究の目的

本研究では，k-means 法を拡張した手法が文書数の少ないトピックを抽出できないことと，異なるトピックにおいて文書の類似度が高い場合に適切にトピックを抽出できないこと，そして，時系列的な特性に基づく手法が文書集合から適切にトピックを抽出できないことの 3 つの問題に対応したトピック抽出手法を新たに提案し，日々，時々刻々と増加する Web ニュースをトピックごとに自動的に整理分類を実現することを目的とする．3 つの問題に対する対応方法を次に示す．

2.1.1 k-means 法の問題に対する対応策

従来の k-means 法を拡張してトピックを抽出する研究[10),11)]では，文書中の単語の出現頻度を基に重要な単語を抽出し，それらを用いて文書を高次元ベクトルで表現した文書ベクトルを用いている．しかし，文書数の少ないトピックの重要語が，他の文書数の多いトピックの出現単語と重複した場合，出現単語の特徴が消失し，これらのトピックに属する文書が文書数の多いトピックとして抽出される．そのため，文書数の少ないトピックを適切に抽出できないという問題が生じる．さらに，k-means 法は，文書間の類似度に依存した分類であることから，トピックを構成する単語の出現傾向が類似している場合，異なるトピックに属する文書を同じトピックとする誤判定が発生する．そのため，文書間の類似度に基づいた分類では，適切にトピックを抽出できないという問題が生じる．そこで，本提案手法では，トピックは任意期間に集中して発生する点を考慮して，バースト期間が類似するバースト語の組合せを遺伝的アルゴリズム[18)]によって多数選出し，多様なトピックの候補を生成することで対応する．本提案手法では，バースト語の組合せ抽出を行う場合に次の条件を考慮する．

- 条件：各バースト語間のバースト期間の始点から終点までの類似度が高い組合せを優先して抽出する．

この条件を設定することで，任意の期間に発生する何らかのトピックと関係があり，かつ時系列情報の関連の強いバースト語の組合せを多数生成することができる．これにより，トピックにおける文書数の多少にかかわらず，任意期間に頻出した単語からのトピック抽出が可能となる．そして，次項（2.1.2 項）の手法によって，トピックの候補として多数選出したバースト語の組合せを用いることにより，文書集合から洩れなくトピック抽出することが可能となる．

2.1.2　Kleinberg の手法の問題に対する対応策

従来の Kleinberg の手法を用いたトピック抽出では，単一のバースト語を含む文書集合をトピックとするため，バースト語どうしの関連は考慮されず，複数のバースト語が同様の文書群を抽出するという問題が発生する．また，1 つのバースト語が複数のトピックに存在する場合，異なるトピックにおいて重複した文書が増加する．さらに，任意の期間において複数のトピックが存在し，多数のバースト語が抽出される場合，各バースト語がどのトピックと密接に関連しているかを特定できない．これらの問題点により，Kleinberg の手法では，文書集合から適切にトピックを抽出できないという問題がある．そこで，本提案手法では，上述の問題に対して，トピック候補であるバースト語の組合せと文書の関係が準最適となる結果を遺伝的アルゴリズムによって選出することで対応する．本提案手法では，トピック抽出を行う場合に次の 2 つの条件を考慮する．

- 条件 1：文書集合全体から洩れなく文書を抽出する．
- 条件 2：トピック間の文章の重複を少なくする．

これらの 2 つの条件を設定することで，文書集合から洩れなくトピックに関連する文書を抽出することができ，かつトピック間の文書の重複が少ない独立したトピックを抽出することが可能となる．

2.2　処理の流れ

本提案システムは，**図 1** に示すバースト検出処理，バースト語の組合せ抽出処理，そして，トピック抽出処理の 3 つの処理によって Web ニュースより時系列的な特性を考慮したトピック抽出を行う．処理の手順は次のとおりである．

バースト検出処理（**図 2** (1)）では，Kleinberg の手法と同様，文書集合からすべての単語を抽出し，各単語が出現する文書を抜き出して時系列順に並べ替える．ここで，任意の単語を含む時系列順に並べ替えた文書集合を時系列文書と定義する．そして，すべての時系列文書に Kleinberg の手法を適用することでバースト語を検出する．

バースト語の組合せ抽出処理（図 2 (2)）では，トピックを構成すると考えられるバースト語の組合せを生成する．本提案手法は，バースト検出処理で抽出した時系列情報であるバースト語のバースト期間に着目する．ここで，バースト期間とは，任意のバースト語が連続して出現する期間を表す．バースト語の中には,「殺人」や「事件」のように 29 日と 30 日の期間で重複するバースト語や,「夏」や「気温」のように 26 日から 31 日までの連続した期間で重複するバースト語も存在する．そこで，本提案手法では，複数のバースト語のバースト期間の類似度を求め，任意のトピックと関連するバースト語の組合せを遺伝的アルゴリズムにより多数生成する．

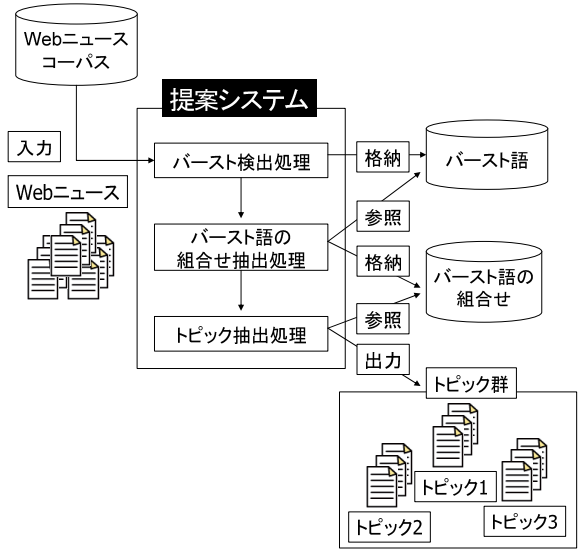

図 1　処理の流れ
Fig. 1　Flow of process.

トピック抽出処理（図 2 (3)）では，上述の処理で生成したバースト語の組合せを利用して，バースト語と文書の関係を考慮したトピック抽出を行う．本提案手法では，バースト語の組合せと各トピックとの関係を選出するのに膨大な計算量を要するため，遺伝的アルゴリズムを用いて準最適解を求める．遺伝的アルゴリズムでは，文書集合の網羅率とトピック間の重複文書の少なさを評価値とする．そして，文書集合から抽出漏れが少なく，かつ各バースト語の組合せから抽出した文書集合間の重複文書が少なくなる最適なバースト語の組合せを抽出する．

本論文では，まず，3 章で既存手法の 3 つの問題を解決する Web ニュースからのトピック抽出について解説する．次に，4 章で本提案手法の有用性を検証するために実証実験を行う．最後に，5 章で本研究により得られた結果についての考察と今後の発展について述べる．

3.　Web ニュースのトピック抽出

本章では，ある期間において頻出する重要な単語を特定するために，まず，Kleinberg の

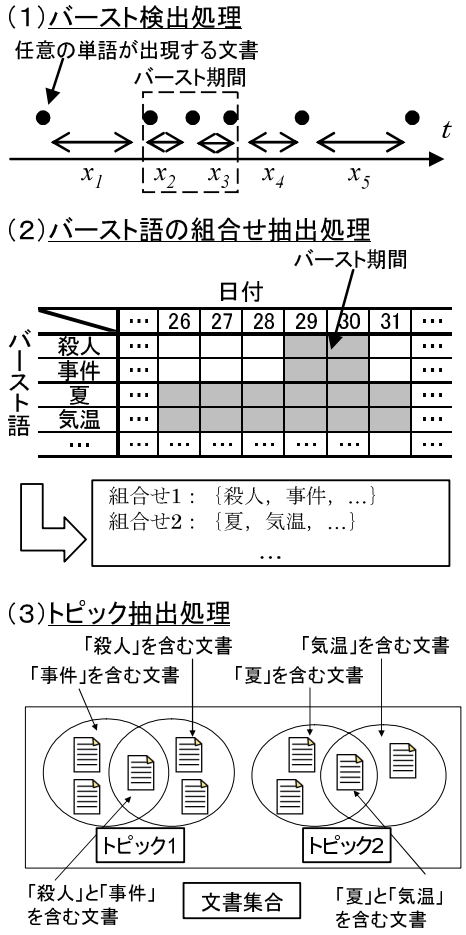

図 2　各処理の概要
Fig. 2　Outline of each process.

手法を利用して文書集合のすべての単語にバースト検出処理を行う．次に，任意の期間においてバースト期間と出現傾向が類似しているバースト語の組合せを生成する．最後に，バースト語の組合せを利用して，文書集合に含まれるトピックを特定し，各トピックに文書を分類する．ここで，バースト検出処理では，文書集合中に含まれるすべての単語について，トピックを表現する適切な単語であるかを評価できるため，Kleinberg の手法を用いた．また，トピック抽出の手法として提案しているバースト語の組合せ，およびそれらの組合せを利用したトピック抽出は，本研究において新たに定義した手法である．次に各処理の詳細を説明する．

3.1　バースト検出処理

近年さかんに行われている時系列データからトピックとなる単語を抽出する研究[12)–17)]では，短期間に頻繁に出現する時系列情報に着目してバースト語を検出する．本研究では，これらの研究においてよく用いられる Kleinberg の手法[13)] を用いて最新のトピックとなりうるバースト語を抽出する．Kleinberg の手法では，時系列情報とある単語を含む文書の出現頻度から，ある単語が頻出している期間とその活性度を算出することができる．Kleinberg の手法を次に示す．

$$f(x_i) = \lambda^{-\lambda x_i} \tag{1}$$

$$\lambda_0 = \left(\frac{P}{T}\right) \tag{2}$$

$$\lambda_1 = S\lambda_0 \tag{3}$$

$$C_j(t) = -lnf_j(x_t) + min_l(C_l(t-1) + \gamma(l,j)) \tag{4}$$

まず，文書が到着する時間間隔 x_i（$i = 1, 2, ..., P-1$），指数分布のパラメータ λ から式 (1) に示す確率密度関数 $f(x_i)$ を求める．ここで，指数分布とは，確率密度関数が式 (1) で表される分布であり，**図 3** に示すグラフで表される．次に，時間区間 $(O, T]$ における任意の単語が出現する全文書数 P を時間区間 T で除算することで，式 (2) に示す文書の到着時間間隔の指数分布のパラメータ λ_0 を求める．式 (2) は単位時間あたりの文書数と定義できる．そして，ユーザ定義の定数 S（$S > 1$）を用いて式 (3) に示す指数分布のパラメータ λ_1 を求める．ここで，定数 S の値が 1 に近いほどバースト語と判定しやすく，値が大きいほどバースト語と判定され難くなる．

図 3 に示す 2 つのグラフ $f_0(x)$ と $f_1(x)$ の交点 Z よりも x の値が大きい場合は $f_0(x) > f_1(x)$ となり，小さい場合は $f_0(x) < f_1(x)$ となる．2 つのグラフの指数分布は，λ_0 を平常状態のパラメータ，λ_1 をバースト状態のパラメータとすると，ある x_t が $f_0(x_t) < f_1(x_t)$ を満たす場合，λ_1 による確率密度関数の値が λ_0 による確率密度関数の値より大きくなり，バースト状態である可能性が高い．同様に，ある x_t が $f_0(x_t) > f_1(x_t)$ を満たす場合，平常状態である可能性が高い．そのため，2 つの確率密度関数 $f_0(x_t)$ と $f_1(x_t)$ の結果を比較することで，文書がバースト状態か平常状態のどちらであるかを判断すること

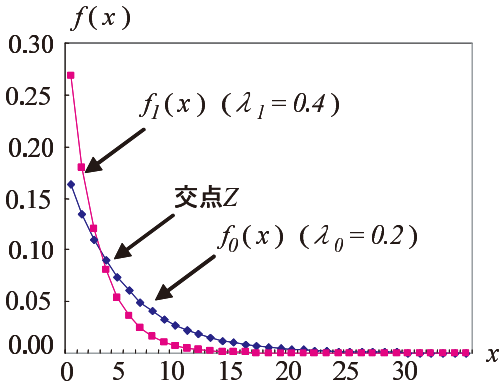

図 3　文書の到着時間間隔の指数分布
Fig. 3　Exponential distribution of document arrival time intervals.

が可能である．

しかし，確率密度関数の大きさを比較するだけでは，バースト状態と平常状態が容易に変化するという問題が発生する．そこで，時系列順に並べ替えた文書に付与した番号 t，文書番号 t と文書番号 $t-1$ の到着時間間隔 x_t，そして状態遷移コスト $\gamma(l,j)$ から式 (4) に示す状態 j（$j=0$，1）であるために必要なコスト $C_j(t)$ を求める．ここで，状態 j は 0 を平常状態，1 を単語が頻出しているバースト状態であることを表す．また，状態遷移コスト $\gamma(l,j)$ における l は，文書番号 $t-1$ が平常状態を表す 0 の状態か，バースト状態を表す 1 の状態のどちらであるかを表す．状態 j であるために必要なコスト $C_j(t)$ では，Viterbi アルゴリズム[20] によって最後の文書 t_{last} からたどり，最小のコスト状態列を形成することで，細かなバーストの発生を防ぎ，まとまったバーストの検出が可能となる．また，平常状態からバースト状態への遷移を容易に発生させないようにすることもできる．

初期状態 $t=0$，$C_0(t)=0$，$C_1(t)=\infty$ として，式 (4) を用いて必要なコストを計算する．状態遷移コスト $\gamma(l,j)$ は，$l<j$ の場合に $\gamma(l,j)=\gamma\log P$，それ以外の場合は 0 とする．ここで，γ（$\gamma>0$）はユーザ定義の定数であり，状態遷移を調整することができる．

次に，本研究では，各単語のバースト期間の類似度に着目した組合せを作成するため，一定期間ごとのバースト回数を算出する．バースト回数を算出するための式として，式 (5) を定義する．

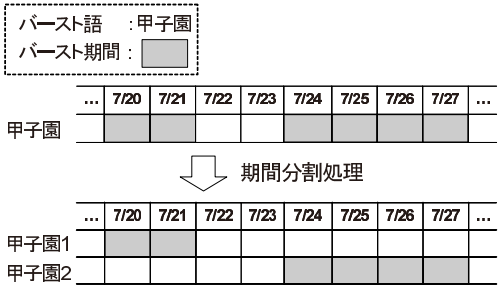

図 4　バースト期間の分割
Fig. 4　Separation of burst period.

$$c_p=\sum_{t=1}^{m}(C(t))\{t\in p\} \tag{5}$$

時間 T を m 個の区間に等しく分割する場合，式 (5) の期間 $p(p=1,2,...,m)$ におけるバースト回数 c_p は，$C(t)=1$ となる t の個数である．本研究では，文書中に出現するすべての単語に式 (4) と式 (5) を適用し，バースト語 $B=(b_1,b_2,...,b_n)$ および各バースト語の期間ごとのバースト回数を算出する．

3.2　バースト語の組合せ抽出処理

バースト語の組合せ抽出処理では，遺伝的アルゴリズムを用いてトピックを構成するバースト語の組合せを抽出する．バースト語の組合せの抽出の前処理として，抽出したバースト語についてバースト期間の細分化を行う．バースト期間の細分化では，同じバースト語が非連続な複数の期間にバーストする場合に別のトピックを表す単語として扱うために，バースト語のバースト期間がつねに連続するようバースト期間を分割する．バースト期間の分割の概念を**図 4** に示す．まず，期間 1 から期間 m までのバースト語 b_e のバースト回数を式 (5) から求める．このとき，b_e のバースト回数を $\{c_{e1},c_{e2},...,c_{em}\}$ とする．次に，バースト回数が 0 となる期間を g とする場合，b_e を b_{e1} と b_{e2} に分割し，バースト語集合 B に加える．分割後の b_{e1} および b_{e2} のバースト回数を式 (6) と式 (7) に示す．

$$b_{e1}=\{c_{e1},c_{e2},...,c_{eg},0,0,...\} \tag{6}$$

$$b_{e2}=\{0,0,...,0,c_{e(g+1)},c_{e(g+2)},...,c_{em}\} \tag{7}$$

式 (6) は b_e の期間 1 から期間 g におけるバースト回数が，式 (7) は b_e の期間 1 から期間 g におけるバースト回数がそれぞれ分割後のバースト語 b_{e1} と b_{e2} に設定されていることを

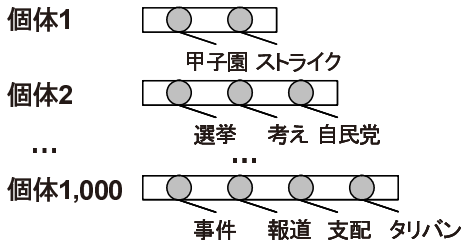

図 5　バースト語の組合せ
Fig. 5　Combination of burst words.

表す．

分割後のバースト語に複数のバースト期間が含まれる場合には，再度バースト語の分割を行い，すべてのバースト語に複数のバースト期間が含まれなくなるまでバースト語の分割を行う．そして，分割したバースト語を利用して，遺伝的アルゴリズムの個体となるバースト語の組合せを抽出する．

本処理の遺伝的アルゴリズムにおける個体を**図 5** に示す．まず，無作為に選出したバースト語の組合せによる個体を作成する．次に，各バースト語のバースト期間が類似する準最適なバースト語の組合せを抽出する．個体が使用するバースト語を遺伝子とし，遺伝子の数は可変となるよう設計する．そして，個体の適応度は，バースト語の任意のバースト期間における出現頻度の類似度を用いて算出する．任意の個体 $Ibursts$ が q 個のバースト語 $B_{Ibursts} = \{b_{Ibursts1}, b_{Ibursts2}, ..., b_{Iburstsq}\}$ を使用する場合，個体 $Ibursts$ の適応度を算出するための式を次のように定義する．

$$MaxPeriod(\alpha, \beta) = Later(L_{\alpha end}, L_{\beta end}) - Earlier(L_{\alpha start}, L_{\beta start}) \quad (8)$$

$$MinPeriod(\alpha, \beta) = Earlier(L_{\alpha end}, L_{\beta end}) - Later(L_{\alpha start}, L_{\beta start}) \quad (9)$$

$$s(\alpha, \beta) = \frac{MinPeriod(\alpha, \beta)}{MaxPeriod(\alpha, \beta)} \qquad \{0 \leq s(\alpha, \beta) \leq 1\} \quad (10)$$

$$Fit_1(Ibursts) = \frac{\sum_{m=1}^{q-1} \sum_{n=2}^{q} s(b_{Iburstsm}, b_{Iburstsn})}{{}_qC_2} \quad (11)$$

式 (8)，式 (9)，式 (10) の概要を**図 6** に示す．任意のバースト語 α，β の時間区間 $(O, T]$ におけるバースト期間をそれぞれ，$L_\alpha = \{L_{\alpha start}, L_{\alpha end}\}$，$L_\beta = \{L_{\beta start}, L_{\beta end}\}$ とする場合，式 (8) は，2 つのバースト語のバースト期間の中で，最も古いバースト発生時間から最も新しいバースト発生時間までのバースト期間の長さを表す．式 (9) は，2 つのバースト語のバースト期間の中で，重複したバースト期間の長さを表す．ここで，$MinPeriod(\alpha, \beta)$ の値が負となる場合は，最小値として 0 を利用するものとする．式 (10) は，互いのバースト語のバースト期間がどれだけ類似しているかを表す．図 6 では，バースト語 α のバースト期間が 7 月 21 日から 7 月 24 日まで，バースト語 β のバースト期間が 7 月 22 日から 7 月 24 日までなので，$MaxPeriod(\alpha, \beta)$ が 7 月 21 日から 7 月 24 日の 4 日，$MinPeriod(\alpha, \beta)$ が 7 月 22 日から 7 月 23 日の 2 日となり，$s(\alpha, \beta)$ が 0.5 となる．式 (11) では，個体 $Ibursts$ が使用するバースト語のすべての組合せのバースト期間の類似度 $\{s(1,2), ..., s(1,q), s(2,3), ..., s(2,q), ..., s(q-1,q)\}$ の平均値を個体の適応度として算出する．この処理により，バースト期間が類似したバースト語の組合せを抽出できる．

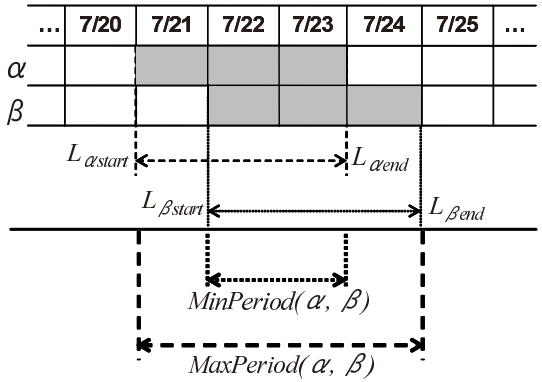

図 6　バースト期間の類似度
Fig. 6　Similarity of two burst periods.

3.3　トピック抽出処理

トピック抽出処理では，バースト語の組合せ抽出処理で作成したバースト語の組合せを無作為に複数選出して個体を作成する．そして，各々のバースト語の組合せを含む文書集合を生成し，文書集合間の文書の重複が少なく，かつ文書集合全体から文書の抽出漏れの少ない準最適なバースト語群の組合せを抽出する．本処理の遺伝的アルゴリズムにおける個体を**図 7** に示す．本処理では，バースト語群の組合せを個体として使用する．個体が使用するバースト語の組合せを遺伝子とし，遺伝子の数は可変となるよう設計する．そして，本手法における個体の適応度は，(1) 個体が使用する個々のバースト語の組合せを含む文書集合と，他のバースト語の組合せを含む文書集合の重複度合い，および (2) 個体が使用するバー

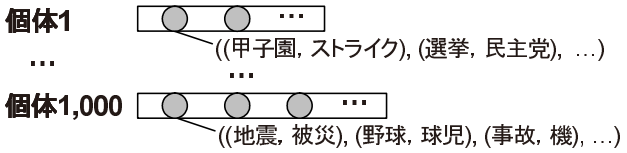

図 7　各トピックのバースト語の組合せ
Fig. 7　Each topic's combination of burst words.

スト語の組合せが関連する文書数の全体文書数に対する網羅率を用いて算出する．重複度合いと網羅率およびその 2 つの指標からバースト語群の組合せである個体 $Igroups$ の適応度を算出するための式を次のように定義する．

$$Dup(A_{Iburstm}, A_{Iburstn}) = \frac{count(A_{Iburstm} \cap A_{Iburstn})}{count(A_{Iburstm} \cup A_{Iburstn})} \tag{12}$$

$$Dup(A_{Igroups}) = \frac{\sum_{m=1}^{y-1} \sum_{n=2}^{y} Dup(A_{Iburstm}, A_{Iburstn})}{{}_yC_2} \tag{13}$$

$$Ind(A_{Igroups}) = 1 - Dup(A_{Igroups}) \tag{14}$$

$$Cov(A_{Igroups}) = \frac{count(A_{Iburst1} \cup A_{Iburst2} \cup ... \cup A_{Ibursty})}{count(A)} \tag{15}$$

$$Fit_2(Igroups) = \frac{e \bullet Ind(A_{Igroups}) + (1-e) \bullet Cov(A_{Igroups})}{2} \tag{16}$$

全文書集合 A に含まれる任意の文書 a が含む単語の集合を $W_a = \{w_1, w_2, ..., w_d\}$ とする場合，$Ibursts \subset W_a$ となるような a の集合を $A_{Ibursts}$ とする．バースト語群 $Ibursts$ と $A_{Ibursts}$ の関係を**図 8** に示す．図 8 では，文書 a_1 に含まれる単語が $\{w_1, w_2, w_4, w_5\}$，文書 a_2 に含まれる単語が $\{w_2, w_3, w_4, w_6\}$，文書 a_3 に含まれる単語が $\{w_4, w_5, w_6, w_7\}$，個体 $Ibursts$ に含まれる単語が $\{w_4, w_5\}$ である．このとき，単語 w_4 と w_5 の両方を含む文書は a_1 と a_3 となる．また，y 個のバースト語群 $Ibursts$ の集合を個体 $Igroups$ とする場合，個々のバースト語群を本文に含む文書集合は，$A_{Igroups} = \{A_{Iburstsl}, A_{Iburstsl}, ..., A_{Iburstsy}\}$ となる．このとき，集合 $A_{Iburstsm}$ と集合 $A_{Iburstsn}$ の両方に出現する文書の集合は $(A_{Iburstsm} \cap A_{Iburstsn})$ で表され，集合 $A_{Iburstsm}$ と集合 $A_{Iburstsn}$ のどちらかに出現する文書の集合は $(A_{Iburstsm} \cap A_{Iburstsn})$ で表される．式 (12) は，集合 $A_{Iburstsm}$ と集合 $A_{Iburstsn}$ の重複率 $Dup(A_{Iburstsm}, A_{Iburstsn})$ を表す．また，式 (13) は，集合 $A_{Igroups}$ に含まれるすべての文書の重複率 $Dup(A_{Igroups})$ を表す．本手法では，個体 $Igroups$ によって全文書集合を複数の文書集合に分類し，分類された文書集合どうしの独立度がより高い個体を遺伝的アルゴリズムによって選択する．文書集合どうしの独立度は，式 (13) で算出した $Dup(A_{Igroups})$ を用いて式 (14) で算出する．式 (14) は，集合 $A_{Igroups}$ に含まれるすべての文書の独立度 $Ind(A_{Igroups})$ を表す．また，式 (15) は，全文書 A に占める $A_{Igroups}$ の割合として網羅率 $Cov(A_{Igroups})$ を表している．式 (15) の概要を**図 9** に示す．図 9 において，分母は文書集合全体を指す．また，分子は $A_{Igroups}$ の要素数を指す．このとき，$Igroups$ の適応度 $Fit_2(Igroups)$ は，式 (16) を用いて $Ind(A_{Igroups})$ と $Cov(A_{Igroups})$ の

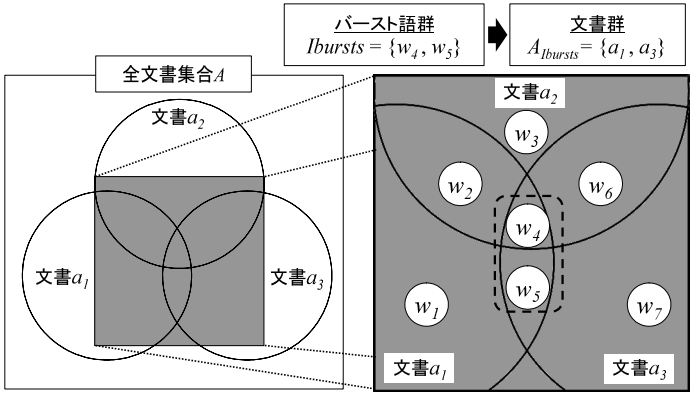

図 8　$A_{Ibursts}$ と $Ibursts$ の関係
Fig. 8　Relation between $A_{Ibursts}$ and $Ibursts$.

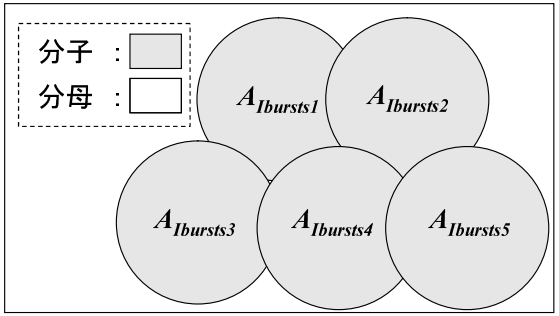

図 9　文書の網羅率
Fig. 9　Coverage of documents.

2 つの指標から算出される．式 (16) の変数 e は，文書の重複度合いと文書の網羅率の重み付け変数として使用する．

この処理により，文書の重複度合いが少なく，文書の網羅率の高い各トピックのバースト語の組合せを抽出でき，抽出したバースト語の組合せによって各トピックに文書の分類を行うことができる．

4. 評価実験

本提案手法の有用性を実証するために，学習データを用いずにトピックを抽出する手法に関する研究で一般的に用いられる k-means 法と Kleinberg の手法の 2 つのトピック抽出手法と比較実験を行い，本提案手法の有用性を確認する．

4.1 実験内容

本実験では，Web ニュースから抽出した任意の時期に出現する複数のトピック抽出結果についての評価を行う．次に比較対象となるトピック群を示す．

- `k-means` 法により抽出したトピック群
- `Kleinberg` の手法により抽出したトピック群
- 提案手法により自動で抽出したトピック群

ここで，k-means 法は，抽出するトピックの数を 10 個に設定してトピック抽出を行った．また，Kleinberg の手法については，抽出したバースト語上位 100 語の中から最も優れた組合せを人手で選出し，その単語を用いてトピックの抽出を行った．

トピック抽出結果の比較では，それぞれの手法で抽出した文書数，トピック数，正解分類数，誤分類数とその抽出結果を評価する．ここで，正解分類数は各手法で抽出したトピックにおける正解文書数，誤分類数は各手法で抽出したトピックにおける誤分類文書数の平均値をそれぞれ表す．また，抽出結果の評価には F 値を用いる．ここで，F 値は再現率と適合率の調和平均であるため，再現率か適合率のどちらかの値が著しく低下すると，その影響を受けて低い値となる．そのため，再現率と適合率の両方の値が高いことが望ましい．本実験に利用する文書は，2007 年 8 月 23 日時点において，新聞記事データベース「ヨミダス文書館[19]」が分類した注目トピック 10 個（文書数 857 件）とした．これらの文書は，注目トピックと関係のある 2007 年 6 月 2 日から 2007 年 8 月 23 日までの Web ニュースで構成されている．ヨミダス文書館から抽出した文書群の内訳を**表 1** に示す．

本実験のバースト語検出に用いた Kleinberg の手法[13] には，2 つのユーザ定義のパラメータ S と γ がある．本実験では，それぞれ一般的に用いられる値[14] として $S = 2$ と $\gamma = 1$ を設定する．また，バースト語の組合せ抽出処理，およびトピック抽出処理の遺伝的アルゴリズムは経験的に**表 2** に示すパラメータを用いた．そして，トピック抽出処理で用いた遺伝的アルゴリズムの評価関数において文書の重複度合いと文書の網羅率の重み付けを同様とするため $e = 0.5$ とした．

表 1　新聞記事データベースから抽出した文書群の内訳
Table 1　Details of documents extracted from news paper database.

トピック	期間	文書数
中国産食品の安全性	2007/6/02 – 2007/8/23	39
アフガン人質	2007/7/21 – 2007/8/23	60
殺人・最近の事件と裁判	2007/7/23 – 2007/8/23	150
世界同時株安	2007/7/25 – 2007/8/23	55
朝青龍処分	2007/7/26 – 2007/8/23	47
参院逆転	2007/7/30 – 2007/8/23	116
夏の甲子園	2007/8/15 – 2007/8/23	150
白い恋人賞味期限改ざん	2007/8/15 – 2007/8/23	44
中越沖地震	2007/8/16 – 2007/8/23	150
中華航空機事故	2007/8/20 – 2007/8/23	46
合計		857

表 2　遺伝的アルゴリズムのパラメータ
Table 2　Parameters of genetic algorithm.

	バースト語の組合せ抽出処理	トピック抽出処理
初期個体数	1,000	1,000
個体	バースト語の組合せ	各トピックのバースト語の組合せ
世代数	1,000	1,000
交叉率	100%	100%
突然変異率	5%	5%
選択手法	エリート戦略 40% ルーレット戦略 60%	
遺伝子	バースト語	バースト語の組合せ
遺伝子長	バースト語数	バースト語の組合せ数

4.2 トピック抽出結果に関する結果と考察

表 1 の実験データを入力として，k-means 法，Kleinberg の手法，および本提案手法をもとに比較実験を行った．k-means 法，Kleinberg の手法と本提案手法の抽出結果を**表 3** に示す．ただし，適用した手法により抽出できたトピック数が異なるため，表 3 に示す正解

表 3　k-means 法，Kleinberg の手法と提案手法の抽出結果
Table 3　Extracting result with the k-means method, the Kleinberg's method and the proposed method.

	k-means 法	Kleinberg の手法	提案手法
抽出文書数	664	521	741
トピック数	7	4	9
正解分類数	75.14	130.25	82.33
誤分類数	19.71	23.50	16.78
再現率	0.82	0.92	0.90
適合率	0.80	0.86	0.79
F 値	0.81	0.89	0.84

分類数，語分類数，再現率，適合率と F 値の平均値算出に用いた分母は異なる．表 3 の結果を分析することにより，他の手法と比較して，本提案手法は次の 3 点で優れていることが分かる．

1 つ目に k-means 法で抽出された文書は 664 件，Kleinberg の手法で抽出された文書は 521 件であるのに対して，本提案手法では 741 件の文書を抽出することができた．この結果から，本提案手法は，857 件の文書集合から抽出した文書数が他の手法よりも多く，約 9 割の文書を抽出できたことが分かる．

2 つ目に各手法で抽出されたトピック数は，k-means 法で 7 つ，Kleinberg の手法で 4 つであるのに対して，本提案手法では 9 つのトピックを抽出できた．この結果から，本提案手法は，より詳細にトピックを抽出できていることが分かる．

3 つ目に本提案手法の正解分類数では，他の手法よりも多くの文書とトピックを抽出しているにもかかわらず，k-means 法よりも優れた値であることが分かる．Kleinberg の手法については，文書数の上位の 4 つのトピックを抽出した結果の平均値であるため，提案手法よりも値が大きくなる結果となった．Kleinberg の手法で抽出できた 4 つのトピックで比較すると，Kleinberg の手法では正解文書数は 130.25 件であるのに対して，本提案手法では正解文書数は 125.75 件を抽出することができた．このことから，本提案手法では，Kleinberg の手法とほぼ同等の正解文書を抽出できた．また，誤分類数を確認すると，本提案手法は他の手法と比較して最も小さい値であることが分かる．これらの結果から，本提案手法は，Kleinberg の手法と同等の正解文書を抽出し，かつ他の手法よりも文書の誤分類数を改善できたことが分かる．

以上の優れた 3 点から，抽出した文書，およびトピックは，他の手法よりも良い結果であることが分かる．また，F 値についても他の手法とほぼ同等の高い値が得られていることから，他の手法と比較して本提案手法は，多くの文章をより詳細に分類できていることが分かる．

4.3　各トピックの抽出精度に関する結果と考察

k-means 法，Kleinberg の手法と本提案手法で抽出されたそれぞれのトピック抽出結果を**表 4** に示す．表 4 では，それぞれのトピックについて，3 つの手法を用いた場合の抽出文書数，適合率，再現率と F 値を示している．これらの結果から，本提案手法は従来手法における次の 3 つの問題点を解決し，優れたトピック抽出結果を得られたことが分かる．

(1)　k-means 法の文書数の少ないトピックを抽出できないという問題を解消

表 4 から,「殺人・最近の事件と裁判」,「夏の甲子園」と「中越沖地震」の 3 つのトピックでは，k-means 法の再現率が 0.57，0.58 と 0.63 と著しく低下していることが分かる．これは，k-means 法の問題点である文書分類において，文書数の少ないトピックを抽出できないことが原因であると考えられる．k-means 法によるトピック抽出結果では，抽出するトピックの数を 10 個と設定したため，分類結果として文書群を 10 個のトピックとして出力していた．しかし，抽出されたトピックの中で,「殺人・最近の事件と裁判」,「夏の甲子園」と「中越沖地震」に関する 3 つのトピックは，2 つずつ抽出されていた．これは，文書数が多いトピックに属する文書が複数のトピックとして誤認識して抽出され，文書数の少ないトピックの文書と混同されたためである．このことが，再現率を著しく低下させた原因であると考えられる．一方，本提案手法では，バースト語と文書の関係を考慮することにより，上述の k-means 法の問題点を解決し，これらの 3 つのトピックにおいて 0.81 以上の高い再現率を得ることができた．

(2)　k-means 法の文書間の類似度に基づく分類では適切にトピック抽出できないという問題を解消

表 4 から,「白い恋人賞味期限改ざん」のトピックでは，k-means 法において，再現率が 1.00 という非常に高い値を得ることができているが，適合率が 0.40 と著しく低下し，結果として F 値が 0.58 と低い値となっている．抽出されている文書群を確認すると,「中国産食品の安全性」の文書が多数含まれており，適合率を下げる原因となっていた．これは，k-means 法において，各トピックの中心文書に対しての類似度に基づいて文書の分類を行っているが，異なるトピックにおいて文書を構成する単語の出現傾向が類似しているため，同様のトピックとする誤判定が発生し，適切にトピックを抽出できないという問題が発生していると考えられる．実際に，2 つのトピックにおいて「食品」や「菌」などの単語が共通して頻出する傾向が見られたため，同じトピックに分類されたと考えられる．一方，本提案手

表 4　各手法を用いて抽出したトピックの詳細
Table 4　Extracting topics using each method.

	k-means 法				Kleinberg の手法					提案手法				
トピック	文書数	再現率	適合率	F 値	バースト語	文書数	再現率	適合率	F 値	バースト語	文書数	再現率	適合率	F 値
中国産食品の安全性	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃	〃
アフガン人質	66	0.98	0.89	0.94	−	−	−	−	−	韓国，誘拐	60	1.00	0.90	0.94
殺人・最近の事件と裁判	108	0.57	0.80	0.67	殺人，殺害	139	0.93	0.78	0.85	殺人，同署	135	0.90	1.00	0.95
世界同時株安	63	1.00	0.87	0.93	〃	〃	〃	〃	〃	融資，サブ，プライム	46	0.84	0.87	0.85
朝青龍処分	〃	〃	〃	〃	〃	〃	〃	〃	〃	停止，巡業	42	0.89	0.49	0.64
参院逆転	124	0.97	0.91	0.94	参院，民主党，自民党	110	0.95	0.95	0.95	自民党，選挙	99	0.66	0.97	0.79
夏の甲子園	100	0.58	0.87	0.70	選手，投手	122	0.81	0.93	0.87	選手，高校	147	0.98	0.77	0.86
白い恋人賞味期限改ざん	109	1.00	0.40	0.58	〃	〃	〃	〃	〃	石屋，製菓	44	1.00	1.00	1.00
中越沖地震	94	0.63	0.85	0.72	被災，刈羽，避難，沖	150	1.00	0.79	0.88	被災，刈羽	122	0.81	0.95	0.88
中華航空機事故	〃	〃	〃	〃	〃	〃	〃	〃	〃	中華航空，本文	46	1.00	0.53	0.70

法で抽出した「白い恋人賞味期限改ざん」のトピック抽出では，F 値が 1.00 という最良の結果を得ることができた．これは，トピック抽出に用いた「石屋」と「製菓」のバースト語が 8 月 15 日から 23 日しか出現せず，かつ他のトピックには出現しないバースト語であったためであると考えられる．ここで,「製菓」は一般名詞,「石屋」は固有名詞であることから，名詞の種類が抽出結果と関係のないことが分かる．このことから，本提案手法は，時系列情報を適用することで，従来手法のように文書における単語の出現傾向に左右され適合率を下げることなく，最良の抽出結果を得ることができることが分かった．

(3)　Kleinberg の手法における文書集合から適切にトピックを抽出できないという問題を解消

表 1 の実験入力データを分析すると，複数の話題が同一期間に存在する事例として，2007 年 7 月 21 日から 2007 年 8 月 23 日の期間（「アフガン人質」,「殺人・最近の事件と裁判」,「世界同時株安」,「朝青龍処分」と「参院逆転」の 5 つのトピックが重複）と，2007 年 8 月 15 日から 2007 年 8 月 23 日の期間（「夏の甲子園」,「白い恋人賞味期限改ざん」,「中越沖地震」と「中華航空機事故」の 4 つのトピックが重複）の 2 つの期間が存在する．これらの期間において，Kleinberg の手法では，各期間 2 件ずつのトピックしか抽出できない結果となった．これは，Kleinberg の手法において複数の話題が同一期間に存在する場合に適切にトピックを抽出できないという問題が顕著に発生したことを示している．Kleinberg の手法において，トピック抽出に用いたバースト語上位 100 語を**表 5** に示す．表 5 を確認すると，トピック抽出に用いた上位 100 件のバースト語のうち，半数以上が共通して出現するため，除外対象となっていることが分かる．これにより，バースト語とトピックの関係を特定することができず，抽出トピック数が低下したと考えられる．一方，本提案手法では，独立性と網羅性の 2 つの指標を用いてバースト語の組合せと抽出文書群の関係を考慮することにより，上述の Kleinberg の手法の問題点を解決し，重複する期間であってもすべてのトピックを抽出することができた．

表 5　Kleinberg の手法で抽出した上位 100 位のバースト語
Table 5　The top 100 burst words extracted by Kleinberg's method.

トピック	バースト語上位 100 語
殺人・最近の事件と裁判	殺人，殺害
参院逆転	参院，民主党，自民党，政権
夏の甲子園	選手，投手，監督，練習，打者，高校，対戦，打線，一死，ナイン，試合，決勝，佐賀，プレー，直球，三塁，安打，声援，二塁，選抜，本塁打，ベスト，勝利，得点，進出，一塁，打球，終了，追加，先発，守備，左腕
中越沖地震	被災，刈羽，避難，沖，新潟，中越，地震
除外したバースト語	写真，調査，対応，夏，チーム，力，生活，柏崎，説明，手，生徒，支援，活動，左，相手，住民，影響，次，事故，発生，水，代表，東，事件，会社，対策，姿，東京，市内，連続，記者，政府，女性，全国，気持ち，右，関係，状況，会長，情報，心，予定，中央，目，表情，外，撮影，会見，記者，声，自分，被害，設置，死，県

5. 今後の発展

本稿では，時系列情報を考慮した遺伝的アルゴリズムによるトピック抽出手法を新たに考案することで，k-means 法を拡張した手法が文書数の少ないトピックを抽出できないという問題と，異なるトピックにおいて文書の類似度が高い場合に適切にトピック抽出できないという問題を解決し，さらに，時系列的な特性に基づく手法が文書集合から適切にトピックを抽出できないという問題を解決した．また，比較実験の結果から，他の非階層分類手法よりも優れたトピック抽出結果を得ることができ，本提案手法の有用性を示すことができた．

今後の課題として，参院逆転,「世界同時株安」と中国産食品の安全性のトピックの結果を分析した結果，次の 2 つの課題点が存在することが分かった．今後は，これらの課題点を解決し，より優れたトピック抽出手法の実現を目指す．

- **遺伝的アルゴリズムにおけるパラメータ設定の課題**

本提案手法において,「参院逆転」と「世界同時株安」のトピックにおいて，従来手法よりも再現率が低い値となった．これは，それぞれのトピックに関する文書を抽出するための適切なバースト語の組合せを選出できなかったため，取得文書数が低下し再現率が著しく低下したと考えられる．この課題を解決するために，本提案手法では経験的に用いた遺伝的アルゴリズムのパラメータに対して，各種パラメータの検証を行い最適なパラメータを設定することで解決できると考えられる．

- **トピックが長期間に分散される場合におけるトピック抽出の課題**

本提案手法において，表 4 を確認すると「中国産食品の安全性」のトピックを抽出できなかった．これは，トピックの出現期間が長期にわたるのに対して文書数が少なく，トピックに関する文書が特定の期間に密集していなかったため,「中国産食品の安全性」に関係するバースト語を抽出できなかったことが原因であると考えられる．この課題に対しては，期間単位に文書群を分割し，その期間ごとにバースト語を抽出する際の閾値 S の最適な値を決定することで，細かなバースト語の組合せの探索が可能となるため，解決できると考えられる．

参　考　文　献

1) 総務省：平成 18 年度版情報通信白書，ぎょうせい (2006).
2) Allan, J.: *Topic Detection and Tracking*, Kluwer Academic Publishers (2002).
3) 高間康史：TDT（Topic Detection and Tracking），知能と情報，日本知能情報ファジィ学会，Vol.17, No.6, p.696 (2005).
4) 岸田和明：文書クラスタリングの技法，Library and Information Science, No.49, pp.33–75, 三田図書館・情報学会 (2003).
5) Wayne, C.L.: Multilingual Topic Detection and Tracking, *Proc. Language Resources and Evaluation Conference 2000*, European Language Resources Association, pp.1487–1494 (2000).
6) Duda, O.R., Hart, E.P. and Stork, G.D.: *Pattern Classification*, Wiley-Interscience (2000).
7) Blei, M.D., Ng, Y.A. and Jordan, I.M.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022, MIT Press (2003).
8) Wang, X. and McCallum, A.: Topics over Time; A Non-Markov Continuous-Time Model of Topic Trends, *The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.424–433, ACM Press (2006).
9) 上田修功，斉藤和巳：多重トピックテキストの確率モデル，情報処理，Vol.45, No.3,

pp.282–289 情報処理学会 (2004).
10) Hatzivassiloglou, V., Gravano, L. and Maganti, A.: An Investigation of Linguistic Features and Clustering Algorithms for Topic Document Clustering, *Proc. 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.224–231, ACM Press (2000).
11) Franz, M., McCarley, J.S., Word, T. and Zhu, W.: Unsupervised and Supervised Clustering for Topic Tracking, *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.310–317, ACM Press (2001).
12) Swan, R. and Allan, J.: Automatic Generation of Overview Timelines, *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.49–56, ACM Press (2000).
13) Kleinberg, J.: Bursty and Hierarchical Structure in Streams, *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1–25, ACM Press (2002).
14) 藤木稔明，南野朋之，鈴木泰裕，奥村　学：documentstream における burst の発見，情報処理学会自然言語処理研究会研究報告，Vol.2004, No.23, pp.85–92 (2004).
15) 崔　春花，北川博之：到着頻度と関連性を考慮した時系列文書の連続的トピック分析，電子情報通信学会データ工学研究会技術研究報告，Vol.104, No.177, pp.19–24 (2004).
16) 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 (2006).
17) 木村昌弘，斉藤和巳，上田修功：Web のトピックダイナミックスモデル，情報処理学会知識と複雑系研究会研究報告，Vol.2004, No.85, pp.155–162 (2004).
18) Holland, H.J.: *Adaptation in Natural and Artificial Systems*, Bradford Books (1975).
19) ヨミダス文書館．https://db.yomiuri.co.jp/bunshokan/
20) 長尾　真：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).

(平成 19 年 10 月 9 日受付)
(平成 20 年 4 月 8 日採録)

中村　健二（学生会員）

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．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）．自然言語処理，知識情報処理，テキストマイニング等の研究に従事．2002 年（株）関西総合情報研究所入社．現在に至る．システム設計，データモデル設計等の研究開発に従事．土木学会学生会員．

田中　成典（正会員）

1963 年生．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，知識情報処理システム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．1997 年助教授．2004 年教授，2006 年関西大学学生センター副所長，現在に至る．博士（工学）．2002 年 8 月から 1 年間カナダの UBC にて客員助教授．専門は知識工学と土木情報学．土木学会，GIS 学会，IABSE，人工知能学会，日本知能情報ファジィ学会と情報知識学会，各会員．1999 年関西経済同友会主催 KSVF ベンチャーアイデア大賞入賞．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同社取締役会長．2006 年（株）フォーラムエイトの顧問に就任．CAD/CG，GIS/GPS，画像処理，そして Web ソリューションビジネスに関連する研究業務に従事．また，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，土木学会土木情報システム委員会土木 CAD 小委員会委員長，土木学会 ISO 対応特別委員会委員，ISO/TC184/SC4 国内委員等を歴任．現在，国土交通省管轄の日本建設情報総合センター建設情報標準化委員会各種委員，オープン CAD フォーマット評議会 OCF 検定監査委員会委員長．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．

古田　　均

1948 年生．1971 年京都大学工学部卒業．1973 年京都大学大学院工学研究科修士課程修了．1976 年同大学院工学研究科博士課程修了．同年京都大学工学部助手．その後講師，助教授を経て，1994 年関西大学総合情報学部教授，現在に至る．その間，米国パディー大学客員助教授，米国プリンストン大学客員研究員，2004〜2005 年米国コロラド大学客員教授．構造物の信頼性解析，最適設計，ライフサイクルコスト解析，ソフトコンピューティングの構造設計・維持管理への応用に関する研究に従事．著書に『ファジィ理論の土木工学への応用』,『建築土木技術者のためのファジィ理論入門』,『遺伝的アルゴリズムの構造工学への応用』,『Life-Cycle Cost Analysis and Design of Civil Infrastructure Systems』等．日本知能情報ファジィ学会，計測自動制御学会，システム制御情報学会，土木学会，日本建築学会，日本材料学会，日本鋼構造協会，ASCE 各会員．

吉村　智史（正会員）

1983 年生．2006 年関西大学総合情報学部卒業．2008 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．修士（情報学）．自然言語処理，知識情報処理，テキストマイニング等の研究に従事．

北野　光一（学生会員）

1983 年生．2006 年関西大学総合情報学部卒業．2008 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）．自然言語処理の研究に従事．2004〜2008 年（株）関西総合情報研究所にて活動．