# 到着頻度と関連性を考慮した文書ストリームのトピック分析

崔 春花† 北川 博之‡

†筑波大学理工学研究科 〒305-8573 つくば市天王台 1-1-1
‡筑波大学電子・情報工学系 〒305-8573 つくば市天王台 1-1-1
E-mail: {hana, kitagawa}@kde.is.tsukuba.ac.jp

**あらまし** 近年ネットワークを介して大量の文書の配信や交換が行われており、それらコンテンツの分析技術の重要性が増加している。重要なコンテンツ分析の１つとして、電子メールやニュース記事などの大規模時系列文書ストリーム中におけるトピック分析がある。本研究では、特に、特定のトピックの時間的な活性度の変化の分析を対象とする。対象とするトピックへ関連性が高い文書が高い頻度で到着するのは、そのトピックの活性度が高い状態であり、そうでない場合には活性度が低い状態と見なす。本論文では、各文書のトピックに対する関連性と到着頻度の両者を考慮した、文書ストリームに対する活性度分析手法を提案する。また、実データを用いた実験によりその有効性を検証する。

**キーワード** 文書ストリーム、トピック分析、トピック活性度、到着頻度，関連性

# Topic Analysis of Document Streams Based on Document Arrival Rate and Document Relevance

Chunhua CUI† Hiroyuki KITAGAWA‡

†Master’s Program in Science and Engineering, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan
‡Institute of Information Sciences and Electronics, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan
E-mail: {hana, kitagawa}@kde.is.tsukuba.ac.jp

**Abstract** Dissemination and exchange of a large amount of documents have become popular according to the advance of network technology in recent years. Thus, importance of content analysis techniques is increasing. Topic analysis in a series of large-scale document streams such as E-mail and news articles is one of such. important research issues. This research especially aims at the analysis of time varying activation levels of topics. When documents of high relevance with a specific topic arrive very frequently, then the activation level of the topic is regarded high, otherwise the activation level is considered to be low. In this research, we propose a systematic topic analysis method for document streams incorporating both document arrival rate and document relevance. Moreover, we evaluate the effectiveness of the proposed method by experiments using real data.

Keyword document stream , topic analysis, relevance, document arrival rate

## 1. はじめに

近年ネットワークを介して大量の文書の配信や交換が急増しつつあり、電子メールやニュース記事などのような時系列的に文書を送信する文書ストリームのコンテンツ分析技術の重要性が増加している。そのようなコンテンツ分析の１つとして、時系列文書ストリーム中のトピック分析がある。これまでに、文書ストリームに含まれるトピックの抽出、出現場所の発見等に関する研究などが行われている［１］［２］［３］。トピック分析の一種として、あるトピックの活性度の分析がある。例えば、ニュース記事文書ストリームにおいてあるトピックに関する記事が頻繁に到着する場合には、一般に、そのトピックの重要性が高かったり世の中の関心が高い状況にあることが多い。このように、対象とするトピックに関連性の高い文書が高い頻度で到着する場合はそのトピックの活性度が高い状態であり、そうでない場合には活性度が低い状態と見なすことができる。このような活性度の分析は、時間軸と関連付けた大規模時系列文書ストリームの構造解析、要約、傾向分析等において極めて重要である。活性度分析においては、文書の内容分析と文書の到着頻度の分析の両者が重要である。既存のトピック分析手法の大

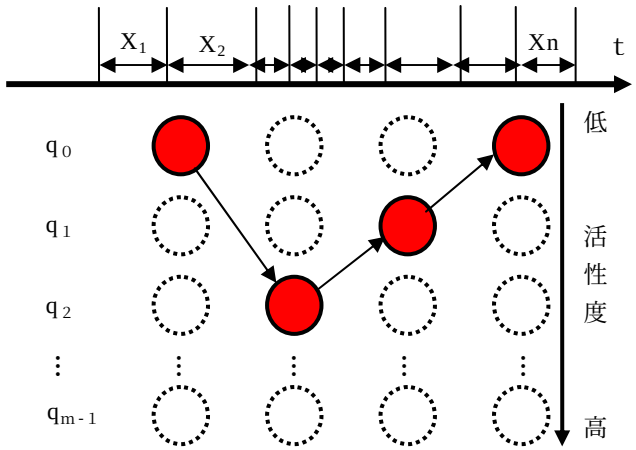
部分は、文書の内容分析を対象としたものであり、文書ストリーム中の文書の到着頻度に着目してトピックを分析する研究は極めて少ない。そのような研究の代表例として Kleinberg による隠れマルコフモデルを用いた分析手法がある［４］。しかし、Kleinberg による分析手法では、逆に文書の到着頻度のみを対象とし、各文書の文書内容が考慮されていないという問題がある。

そこで、本研究では、Kleinberg の分析手法をベースとして、あるトピックに対する文書の関連性と文書の到着頻度の両者を考慮した文書ストリームの活性度分析手法を提案する。また、CNN ニュース記事文書ストリームを対象とした実験により、本提案手法の有効性を確認する。

本論文の２節では、Kleinberg の分析手法の概要を示す。３節では、本論文での提案手法を述べる。４節では CNN ニュース記事文書ストリームを用いた本手法の実験評価を示す。５節では結論と今後の課題を述べる。

## 2. Kleinberg の分析手法

Kleinberg は、文書ストリームにおけるあるトピックに関する文書の到着頻度に着目し、そのトピックの活性度を分析する手法を提案した。（論文［４］においては、文書が非常に頻繁に到着する状態をバースト(burst)と呼ぶ。）Kleinberg による手法では、内部状態に応じて文書の到着時間間隔が確率的に決定される隠れマルコフモデルを用いた分析を行う。隠れマルコフモデルはマルコフモデルの各状態に対して、確率的な記号の出力を加えたモデルである。活性度が高い状態では頻繁に文書が到着するため、到着時間間隔は確率的に短くなり、逆に活性度が低い状態ではまれにしか文書が到着しないため、到着時間間隔は確率的に長くなる。簡単な例を図１に示す。図１では時間軸 $t$ にそった時系列文書の到着を示す。縦線は文書の到着を表し、 $x_0, x_1, \ldots, x_{n-1}$ は文書間の到着時間間隔を表す。Kleinberg の手法では、個々の到着時間間隔 $x_i$ は付随する隠れマルコフモデルの内部状態に応じて確率的に出力される記号であるとみなす。いま、 $q_0$ から $q_{m-1}$ までのｍ個の状態を持つ隠れマルコフモデルを仮定し、$q_0$ から $q_{m-1}$ まで状態番号が増加する程、確率的により短い到着時間間隔を与えるものとする。したがって、この順に活性度は高くなることになる。例えば、図１では最初は文書の到着時間間隔が長いので活性度は低い状態にあると見なせるが、到着時間間隔が短くなった場合には、活性度の高い内部状態への遷移が生じたと見なすことができる。さらに、到着時間間隔が元の長い状態に戻った場合には、内部状態も活性度が低い状態に戻ったと見なすことができる。このように、文書の到着時間間隔を活性度に応じた内部状態遷移に反映することで、時間的な活性度の変化をモデル化するというのが基本的な考え方である。




状態遷移列：{q0,q2,q1,q0}

図１.Kleinberg の分析手法

より具体的には、各状態 $q_i$ において、文書の到着時間間隔は所定の指数分布に基づいて確率的に決定されるものをする。つまり文書 $i$ とそれに続いて到着する文書 $i+1$ の間の時間間隔 $x_i$ は指数確率密度関数 $f_i(x_i) = \alpha_i e^{-\alpha_i x_i}$ によって決定されるものとする。ここで $\alpha_i = (n/T)\beta^i$ ( $n$ は全文書数、$T$ は全時間幅、$\beta$ ＞1はパラメータ)は単位時間当たりの文書の平均到着率であり、その逆数が平均到着時間間隔となる。したがって、$\alpha_i$ が大きい状態ほど頻繁に文書が到着する活性度が高い状態を表す。

一連の文書到着時間間隔列 $x = (x_0, x_1, \ldots, x_{n-1})$ が与えられた時、最適の状態遷移列 $s = (s_0, s_1 \ldots, s_{n-1})$（各 $s_i$ は状態番号を表す）は下記のコスト関数を最小にするものとして求める。

$$c(s \mid x) = \left(\sum_{i=0}^{n-2} \tau(s_i, s_{i+1})\right) + \left(\sum_{i=1}^{n-1} -\ln f_{s_i}(x_i)\right)$$

ここで、この式の第１項は、内部状態が状態 $q_{s_i}$ から $q_{s_{i+1}}$ に遷移する際のコストの総和であり、$s_{i+1}$ が $s_i$ より大きい高い場合（ $q_{s_{i+1}}$ が $q_{s_i}$ より活性度が高い場合）は、遷移コストは状態番号の差に比率するものと

する。すなわち、$\tau(j,k)=(k-j)\gamma \ln n$（$\gamma>0$ は状態変化の容易さをコントロールするパラメータ）としている。また、そうでない場合の遷移コストは0としている。第2項は、各状態 $q_{s_i}$ が到着時間間隔 $x_i$ を発生しやすい程小さいコストとなる。すなわち、上記のコスト関数は各到着時間間隔 $x_i$ を状態遷移の概念で説明する上で適切であると同時に、あまりに頻繁な状態遷移が発生するのを防止するよう設計されたものである。

文書到着時間間隔列 $x=(x_0, x_1,..., x_{n-1})$ に対して上記コスト関数を最小化する状態遷移列 $s=(s_0, s_1..., s_{n-1})$ は、隠れマルコフモデルに対するビタビアルゴリズムを用いて求めることができる。詳細は省略するが、概要は次の通りである。$(x_0, x_1,..., x_i)$ に対応する状態遷移で状態 $q_j$ で終了するものの最小コストを $c_j(i)$ で表す。$c_j(i)$ は、初期状態が $q_0$ にあるものとして、$i$ を順次増加させながら次の式を計算することで求めることができる。

$$c_j(i) = -\ln f_j(x_i) + \min_l (c_l(i-1) + \tau(l, j))$$

## 3. 提案手法

Kleinberg の手法では、あるトピックと関連がある文書をキーワード検索等の何らかの手法で特定した後に、上記の方法でそのトピックの活性度を分析することを想定している。すなわち、個々の文書とトピックの関連の度合いについては、一切考慮していない。同じのトピックに関する文書でも、時期によって当然そのトピックとの関連性が強い文書と弱い文書が存在する。したがって、個々の文書は関連性に応じてそれぞれの重要度を区分することができる。また、各文書が含む情報を分析する上でこのような関連性の度合いを考慮することが重要なことは、これまでの情報検索に関する研究で広く認識されている。そこで、本研究では、文書の到着頻度と個々の文書の関連性の度合いの両者を考慮するよう、Kleinberg の手法を拡張する。

図2の（1）では時間軸上にそって到着する文書の関連性の度合いを縦方向の矢印の長さで表している。すなわち、2つの文書はそれぞれ異なる関連性の度合い $R_i$ と $R_{i+1}$ を持つ。Kleinberg の手法では、図2の（2）に示すように、これらをそれぞれ同じ関連性の度合いを持つと見なしていると考えることができる。関連性の度合いを考慮するよう Kleinberg の手法を拡張する一つの方法は、各状態を文書の到着時間間隔の発生頻度だけで区別するのではなく、生成される文書の関連性の度合いも考慮するように拡張することである。しかし、この場合、解析のためのモデルの複雑さが増加し、解析もより困難になる。そこで、本研究では、各文書の関連性の度合いを考慮して文書の到着時間間隔を補正することを試みる。

一般に、ある時刻に到着した文書の影響力は時間と共に次第に逓減すると考えることができる［5］［6］。この考え方を本研究に当てはめると、トピックとの関連性の度合いが弱い文書が到着した時点での状態は、より強い関連性の度合いをもつ文書が到着してから一定の時間が経過した後の状態と同等と見なすことができる。

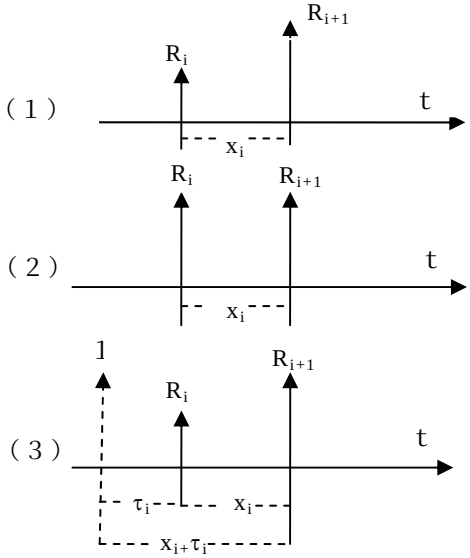



図2．到着時間間隔と関連性の度合い

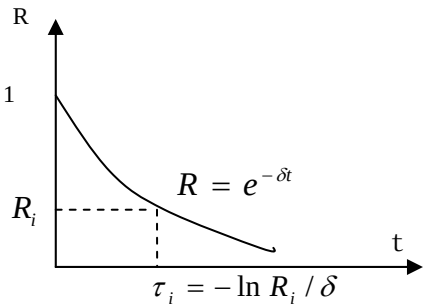



図3.文書の影響力の逓減モデル

具体的な文書や情報の影響力の時間的逓減の割合を表すためのモデルとしては、図3に示すような指数関数逓減モデルがこれまでの研究の中ではしばしば用いられている［5］［6］。すなわち、文書の影響力は時間 $t$ の経過と共に、$R=e^{-\delta t}$ のような関数関係で逓減するとみなす。ここで R は文書の影響力であり、

| 実験１ | キーワード集合 |
|---|---|
| 「Oprah Lawsuit」 | Winfrei, oprah, rancher, cattl, Amarillo, defame, beef, texa, cow, cattlemen |
| 「Bombing　AL Clinic」 | Rudolph, eric, alabama, birmingham, clinic, bomb, truck, lyon, nurs, emili |

表１．実験で実際に使用したキーワード

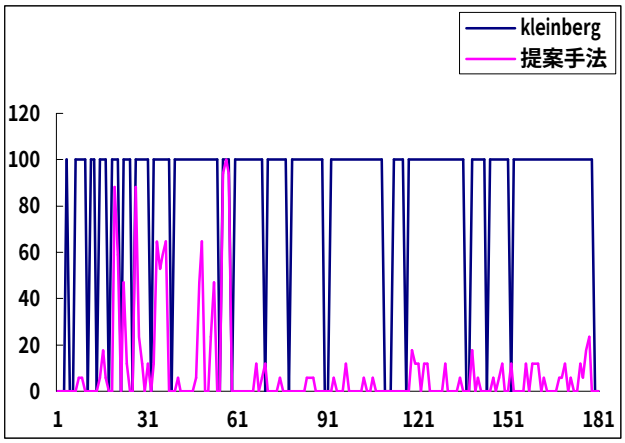



図４．「Oprah Lawsuit」における活性度の変化

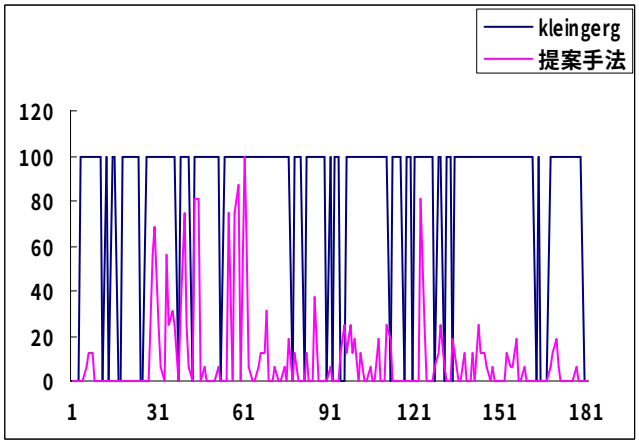



図６．「Bombing　AL Clinic」における活性度の変化

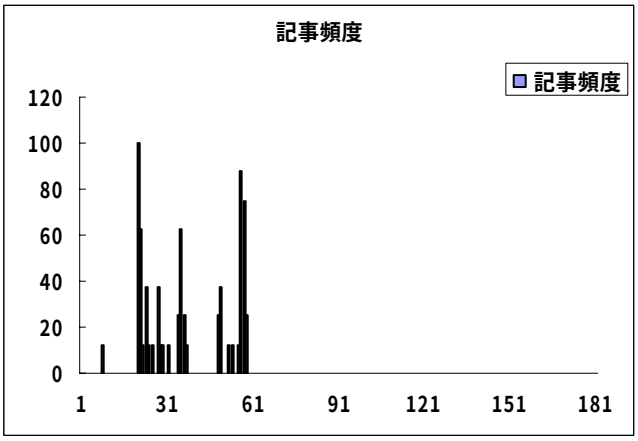



図５．「Oprah Lawsuit」の記事の到着数

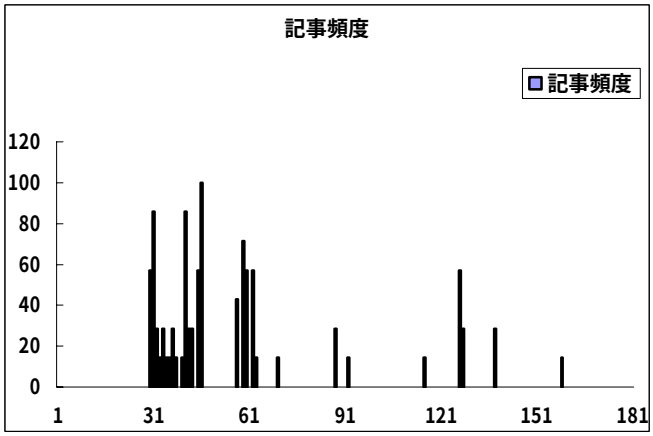



図７．「Bombing　AL Clinic」の記事到着数

δは影響力の時間的減少の割合を決定するパラメタである。本研究では、このモデルを適用する。すなわち、関連性の度合い $R_i$ を持つ文書の時刻 $t$ での到着は強さ１の文書の時刻 $t-\tau_i$ における到着と同等と見なす。より具体的には、図１の（３）に示すように、トピックとの関連性の度合い $R_i$ がの文書が到着してから時間 $x_i$ 経過後に次の文書が到着した場合、関連性の度合いが１の文書が到着してから時間 $x_i + \tau_i$ 経過後に次の文書が到着したものとみなす。ただし、文書の関連性の度合い $R_i$ と時間の $\tau_i$ 間の関係は、上記の指数関数モデルにしたがって、$\tau_i = -\ln R_i / \delta$ とする。すなわち、もともとの到着時間間隔列 $x = (x_0, x_1, ..., x_{n-1})$ は $x' = (x_0 + \tau_0, x_1 + \tau_1, ..., x_{n-1} + \tau_{n-1})$ に変換される。このような変換後の最適状態遷移列は、もともとの Kleinberg モデルと同様の方法で求めることができる。

## 4. 実験

本節では提案手法に対する実験評価を示す。実験１では、提案手法と Kleinberg の手法を用いた実データの分析結果を示し、その比較を行った。次に、実験２では、提案手法において $\beta$、$\gamma$、$\delta$ などの各種パラメタを変化させ、それが実験結果に及ぼす影響を評価した。

実験データとしては、TDT（Topic Detection and Tracing Evaluation）用の評価データの一部である

1998.1.1～1998.6.30 の CNN ニュース記事 21587 件を使用した。これらニュース記事は、様々なトピックのニュース記事を含むが、一部トピックについては、どの記事がそのトピックの関連記事かというラベル付けがあらかじめ人手で与えられている。これらのニュース記事に語幹抽出と不要語除去を行った後のテキストを実験データとして用いた。

### 4.1 実験 1

実験 1 では、提案手法と Kleinberg の手法による活性度の分析結果の比較を行った。具体的な実験方法としては、最初にあるトピックを記述するキーワード集合を１つ与え、全ての記事について関連性の度合いを計算する。本研究では、文書検索で通常行われるように、tf/idf による重み付けを考慮した余弦尺度を用いて与えられたキーワード集合に対する各記事の関連性の度合いを求めた。また、本実験では、関連性の度合いが 0 より大きいすべての記事を対象として分析を行った。

本実験では、CNN ニュース記事中に現れる「Oprah Lawsuit」と「Bombing　AL Clinic」の２種類のトピックを記述するキーワード集合を与えた。これらのトピックを記述するキーワード集合は、次のように求めた。実験データ中の 30 件以上の記事を含む主要なトピック（これらの中には「Oprah Lawsuit」と「Bombing　AL Clinic」が含まれる）20 を選択し、各トピックに属する全ての記事を連結したものを１つの文書とみなす。次に、これらの 20 個のトピックに対応する文書の文書ベクトルを通常の tf/idf を用いて計算する。最後に、各トピックに対する文書ベクトル中で tf/idf の値の大きい単語 10 個を選択する。実験で実際に使用したキーワードを表１に示す。また、実験１では、各パラメタを $\beta$=1.1、$\delta = 5$、$\gamma = 0.01$ としている。

あらかじめ与えられたトピックラベルによると、トピック「Oprah Lawsuit」を表す記事は、1998.1.1 から起算して８日目から５９日目の間に 59 件出現している。また、この期間以外には「Oprah Lawsuit」にラベル付けされた記事は存在しない。また、トピック「Bombing　AL Clinic」を表す記事は、1998.1.1 から起算して２９日目から１５８日目の間に 73 件出現している、また、この期間以外にこのトピックラベルの付けられた記事は存在しない。

図４は、トピック「Oprah Lawsuit」に対応するキーワードを与えた場合の、提案手法と Kleinberg 手法による分析結果を示す。また、図 6 はトピック「Bombing AL Clinic」に対する同様の分析結果である。両グラフにおいて、横軸は 1998.1.1 から起算した日数を表しており、縦軸は状態を表しており、上にいくほど活性度が高い状態を示す。ただし、それぞれの方法で得た活性度の値を最高値が 100 になるように正規化している。また、図５と図７では各トピックのラベルがついた記事がそれぞれの日に何件実際に到着したかを棒グラフで示している。横軸は 1998.1.1 から起算した日数を表しており、縦軸は指定されたトピックラベルがついた記事の到着件数を示す。ただし、ここでも到着件数の最高値が１００となるよう正規化した。トピック「Oprah Lawsuit」に関する実験では、関連性の度合いが０より大きい記事は全部で 858 件抽出された。また、トピック「Bombing　AL Clinic」に関する実験では、９８７件抽出された。図４と図６に示されているように、Kleinberg の手法で得た状態遷移では、全期間に渡って活性度の高い状態が出現している。一方、提案手法では、図 5 や図７に示した記事の到着パターンとの整合性が高い結果が得られている。このように、Kleinberg の手法では、現実の活性度からの乖離が大きい。それに対して、提案手法から得た活性度は多少ノイズと思われる部分が含まれているものの、現実の活性度の変化により近いものとなっている。

### 4.2 実験２

実験２では、提案手法において各種パラメタ値を変化させた場合の分析結果に対する影響を評価した。本実験では、実験１においてトピック「Oprah Lawsuit」を記述するのに用いたキーワード集合を与えた。

#### 4.2.1 パラメタ $\beta$ の変化

２節で説明したように、文書の到着時間間隔は所定の指数分布に基づいて確率的に決定されるものとしている。ここで、指数確率密度関数 $f_i(x_i) = \alpha_i e^{-\alpha_i x_i}$ における $\alpha_i = (n/T)\beta^i$ ( $n$ は全文書数、$T$ は全時間幅、$\beta$ ＞1 はパラメタ)は状態 i における単位時間当たりの文書の平均到着率である。本実験では、$\delta = 5$、$\gamma$ =0.01 とし、$\beta$ を $\beta$＝1.1 および $\beta$=1.5 とした場合の実験結果を比較する。$\beta$ が 1 に近い程状態間の平均到着率の差は小さくなるため、より細かい状態の変化をとらえることが可能となる。図８に実験結果を示す。グラフの横軸と縦軸の意味はこれまでと同様であり、また、活性度の最高値が１００となるよう正規化している。

図８から分かるように、$\beta$＝1.1 の場合の方がより細かい状態変化を表現できている。一方、$\beta$=1.5 の場合は現れる状態の種類が少なく活性度の表現が粗くな

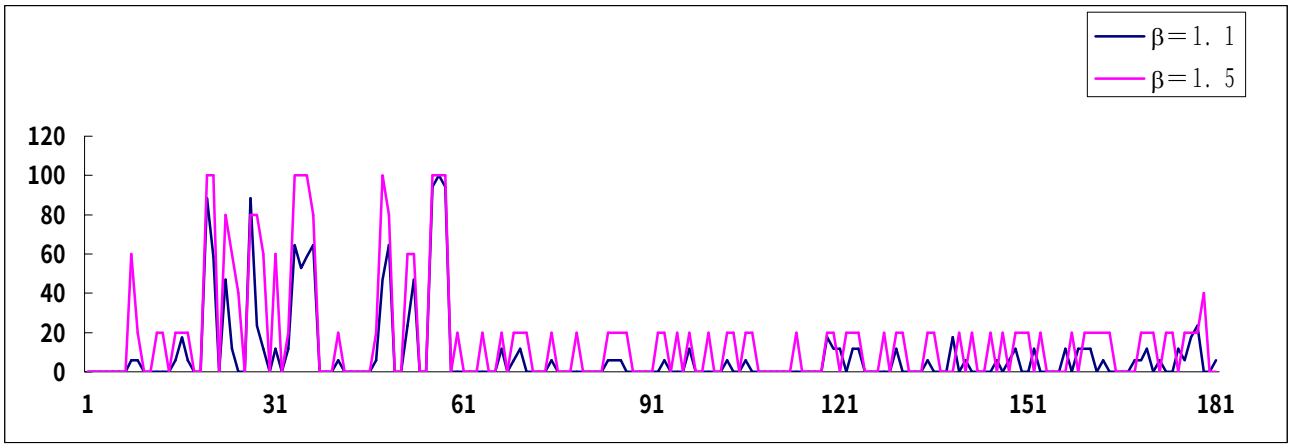



図８．$\beta$の変化による活性度の変化

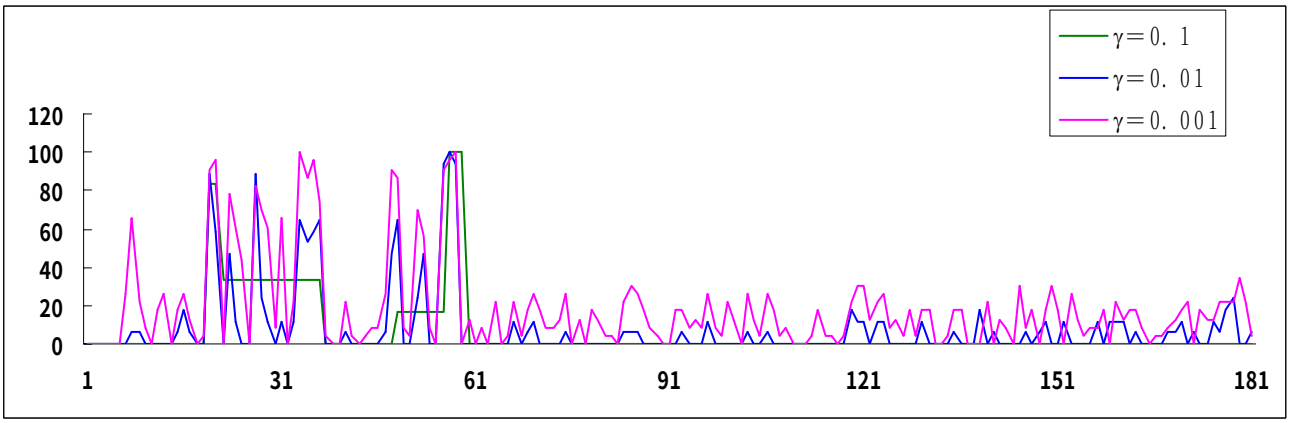



図９．$\gamma$の変化による活性度の変化

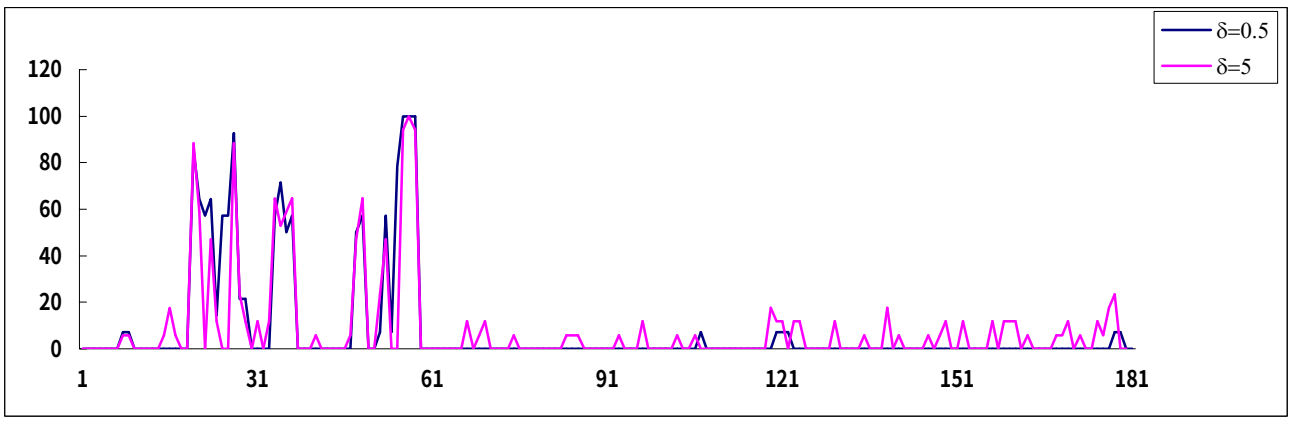



図１０．$\delta$の変化による活性度の変化

っている。ただし、分析対象により適当なきめ細かさの程度は異なるため､それに応じた の選択が必要である。

### 4.2.2 状態遷移コストの変化

与えられた文書到着時間間隔列に対する最適の状態遷移列を決定するためのコスト中には、内部状態遷移コストが含まれる。すなわち、内部状態が状態$q_{s_i}$から$q_{s_{i+1}}$に遷移する際、$s_{i+1}$が$s_i$より大きい場合（$q_{s_{i+1}}$が$q_{s_i}$より活性度が高い場合）は、遷移コストは状態番

号の差に比率するものとする。具体的には、$\tau(j,k) = (k - j)\gamma \ln n$（$\gamma > 0$ は状態変化の容易さをコントロールするパラメタ）としている。本実験ではパラメタ $\gamma$ の変化が実験結果に及ぼす影響を分析する。本実験では、$\delta$ = 5，$\beta$=1.1 とし、$\gamma$ を $\gamma$ =0.1，$\gamma$ =0.01，$\gamma$ =0.001 と変化させた場合の実験結果を比較する。

図９は実験結果である。この図から分かるように、$\gamma$ が小さいほど細かい変動を反映したグラフとなる。一方、$\gamma$ を大きくした場合には細かい変動は少なくなり、特に $\gamma$ =0.1 の場合には顕著な傾向のみしか示さなくなることが分かる。

### 4.2.3 逓減率の変化

提案手法では、情報の影響力の時間的逓減の割合を表すためのモデルとして、文書の影響力は時間 $t$ の経過と共に、$R = e^{-\delta t}$ のような関数関係で逓減するとみなす。ここで R は文書の影響力であり、$\delta$ は影響力の時間的減少の割合を決定するパラメタである。本実験では、$\beta$ = 1.1，$\gamma$ =0.01 として、$\delta$ を $\delta$ = 5 および $\delta$ = 0.5 とした場合を比較する。

図１０に実験結果を示す。このグラフから分かるように、$\delta$ を小さくする程細かい変動を表しにくくなり顕著な傾向のみしか示さなくなることが分かる。なお、本手法において $\delta$ = ∞とした特殊な場合が、Kleinberg 手法に相当する。この意味でも、本手法は Kleinberg 手法の拡張になっている。

## 5. まとめと今後の課題

本研究では、文書の関連性と到着頻度の両者を考慮した大規模時系列文書ストリーム中でのトピックの活性度の分析手法を提案した。本研究では、２種類のトピックを対象に、ＣＮＮニュース記事を用いた実験により、提案手法が従来手法よりも有効であることを証明した。また、提案手法で現れた各種パラメタが状態遷移にもたらす影響を詳細な実験で示した。

今後の課題としては、より詳細な実験検討に加え、文書の到着頻度と関連性の両者を統合的に扱うための他の方法の検討や本提案手法との比較等がある。

## 謝辞

本研究の一部は、文部科学省科学研究費補助金特定領域研究(2)(15017207)、ならびに日本学術振興会科学研究費補助金基盤研究(B)(15300027)による。

## 参考文献

[1]F. Walls, H. Jin, S. Sista, and R. Schwartz, "Topic Detection in Broadcast News", Proc. DARPA Broadcast News Workshop, 1999.

[2]J. M. Schultz and M. Liberman, "Topic Detection and Tracking using idf-Weighted Cosine Coefficient", Proc. DARPA Broadcast News Workshop, 1999.

[3]H. Li and K. Yamanishi, "Topic Analysis using Finite Mixture Model",Information Processing and Management, Vol. 39, 2003.

[4] J. Kleinberg, "Bursty and Hierarchical Structure in Streams", Proc. ACM SIGKDD, 2002.

[5] Y. Ishikawa, Y. Chen, and H. Kitagawa, "An On-Line Document Clustering Method Based on Forgetting Factors", Proc. 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001), September 2001.

[6] B. K. Yi, et al., "Online Data Mining for Co-Evolving Time Sequences", Proc. 16th International Conference on Data Engineering, 2000.