時系列データに意味的に関連するニューストピックの発見

張　一萌† 　何　書勉‡ 　小山　聡‡ 　田島　敬史‡ 　田中　克己‡

†京都大学工学部情報学科　〒606-8501　京都市左京区吉田本町
‡京都大学情報学研究科社会情報学専攻　〒606-8501　京都市左京区吉田本町
E-mail: †‡{zhangym,shumian,oyama,tajima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　本論文では，与えられた時系列データに，意味的に関連のあるニューストピックを発見するシステムを提案する．関連記事を検出する従来の手法の多くが，主に文書の類似度を検出尺度としているのに対して，本提案では，内容類似度のほか，ニュース記事の出現頻度を利用する．これによって時系列データの変動への影響も含むような関連を検出する．具体的な手法として，ニュース記事を話題ごとに分け，次に話題ごとの出現頻度の時間変動と入力された時系列データを比較し，各話題に評価値をつける．さらに，内容類似度によりつけた評価値と合わせて，ある話題が時系列データに意味的な関係の有無を判断する．最後に，話題に関する記事の出現頻度の時間変化の特徴によって，特定の話題がどの時間帯でどのように入力された時系列データに影響を与えていたかを分析する．本論文は，いくつかの時系列データを用いた実験を通して，提案手法の有効性を検証する．

キーワード　Web とインターネット，時系列データ，データマイニング，関連トピック

Discovery of Semantically Related Topics for Given Time Series Data

Yimeng ZHANG† 　Shumian HE‡ 　Satoshi OYAMA‡ 　Keishi TAJIMA‡ 　Katsumi TANAKA‡

† Department of Informatics, Faculty of Enginieering, Kyoto University,
Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8501 Japan
‡ Department of Social Informatics, Graduate School of Informatics, Kyoto University,
Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †‡{zhangym,shumian,oyama,tajima,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　We propose a method for discovering semantically related news topics for given time series data. As compared with most existing methods which find related news mainly based on the similarity among documents, we use both textual and temporal behavior of the news, and expect to find related news which has some impact on the time series data. To this end, we first classify all of the news into events, and then decide whether an event is semantically related to the given time series based on the textual and temporal correlation between the event and the time series data. Finally, we detect when and how a certain event has impacted the time series data, by analyzing the temporal feature of the event. At the end of this paper, we show the properties of our approach by some experiments.

Key words Web and Internet, Time Series Data, Data Mining, Related Topics

1. はじめに

インターネット技術の進歩により，膨大な量のニュースが Web 上で報道されるようになった．最近の調査では，人々がインターネットを利用する主な目的の一つとして，ニュースの閲覧と検索が挙がられている[1].

ニュースサイトについてのアンケート[2]により，インターネットニュースに対する不満は，「一本あたりの記事の情報量が少ない」，「ネット独自の機能や情報が少ない」などが取り上げられている．実際，Yahoo！ニュース[3]などのサイトでは，特定の出来事を報道するだけで，その出来事の関連する情報や起因の分析などが少ない．これに対して，テレビニュースでは，視聴者の注目度の高いニュースを報道する際に，専門家やジャーナリストを招き，ニュースに関するさまざまな情報の分析を行う．例えば，「小泉首相の支持率は過去最高になった」という記事を読むと，小泉首相の支持率に関連あるいは影響しているのはいったい何だったのか，また，どのように影響しているのか（支持率を上昇させたか，低下させたか）を知りたくなる．

ニュースサイトでニュースを報道する際，専門家を招くことはないが，インターネット上の大量のニュース記事に情報技術を利用することにより，自動的に社会的な性質の解明が期待される．本研究では，与えら

れた時系列データ（例えば，内閣支持率や株価など）に意味的に関連のあるトピックを発見する手法を提案する．ここでいうトピックとは，特定の出来事あるいは話題（例えば，郵政民営事業化やライブドア事件など)に関するニュース記事のカテゴリである.ただし，本論文ではトピックに関するニュース記事の出現頻度の時間変動を利用するため，短時間（１週間や２週間）にだけ出現してなくなるようなトピックは扱わない．

従来の「関連記事」とは，内容が類似している記事のことである．関連記事を検出する手法も主に文書の類似度を検出尺度としている．本研究では，入力された時系列データに内容的に関連している記事だけでなく，特に時系列データの変動に影響を及ぼすような出来事に関連する記事を発見しようとする．この意味での関連記事を発見するためには，従来の文書の類似度だけによる手法は不十分だと思われる．そこで，本論文は類似度のほかに，トピックに関するニュース記事の出現頻度の時間変動を利用する手法を提案する．

基本的なアイデアとして，あるトピックに関するニュース記事が多く報道されるたびに入力された時系列データが大きく変動したら，そのトピックは与えられた時系列データの変動に影響を与えていると仮定する．逆に言えば，時系列データの変動に影響があるトピックに対して，多く報道される時間帯で時系列データが変動したら，そのトピックの影響を受けていると仮定する．この仮定のもとで，トピックに関するニュース記事の出現頻度の時間変動と入力された時系列データを比較することによって，トピックは入力時系列データに影響を及ぼしているか，またどの時間帯でどのように変動させたか（上昇させたか，低下させたか）を求めることができる．

本研究の特徴は以下のとおりである．

- ユーザから時系列データとその特徴をつけるキーワード（時系列データの名前）を受け取る．
- 与えられた時系列データに意味的に関連するトピックをランキングして提示する．ここでいう「関連」は二つの意味を含む．
  - 文書の類似度．この類似度は，トピックに属するニュース記事の内容がどれだけ入力された時系列データの内容（時系列データの特徴を示すキーワードによって示されている）に類似するかを示す．
  - 時間的相関．これはトピックが入力時系列データに影響を与えているかどうかを示す．トピックに関するニュース記事の出現頻度の時間変動と与えられた時系列データの相関によって計算される．
- 影響があると判断されたトピックについてどの時間帯で影響を与えたか，またどうやって変動させたか，つまり上昇させたかあるいは低下させたかを判断する．注意すべきことは一つのトピックでも違う時間帯で時系列データを上昇させたり，低下させたりする可能性がある．

以下，2 章では本研究と関連のある研究を紹介し，3 章と 4 章はそれぞれ時系列データに関連するトピックの発見，特定のトピックの時間帯別影響の算出の提案手法について述べる．5 章で提案手法の実装と実験結果について述べる．6 章ではまとめと今後の課題について述べる．

2. 関連研究

ニュース記事を自動的にトピックごとに分ける研究では，TDT（The Topic Detection and Tracking）プロジェクト[4]がある. TDT ではニュース記事をトピックに分類するトピック検出のほか，与えられたトピックの記事を検出するトピック追跡[5][12]に関する研究を行われている．本研究ではトピック検出でできたトピックから入力された時系列データに意味的に関連があるトピックを提示する．また，トピックの動向を把握するため，トピック内のニュース記事の間の内容的な類似，相違の発見を行いながら，記事の時間変動を提示する研究は行われている[6][7]. 本研究ではトピックの時間変動を利用する点で似ているが，ニュース記事の内容の時間変動を一切使わず，ニュース記事の出現頻度の時間変動を使って，それが入力時系列データに意味的に関連があるかどうかを分析する．

社会学や経済学で特定の話題に関するニュース記事が特定の時系列データ（為替レートや支持率）に影響があるかどうか，またどのように影響するかを調べる研究が多くある[8][9][10]. 本研究では特定トピックのニュースが時系列データに影響するかどうかを判断するが，それによって影響があると判断されたトピックを提示するため目的が異なる.そして,本研究では，結果として影響力の分析を自動化するシステムを作り上げるという点もこれらの研究と異なっている．

また，検索エンジンに投げるクエリの類似性を判断するには自然言語処理を利用する代わりにクエリの出現頻度の時間変動を利用する研究がある[11]. この研究では，出現頻度の時間的相関が強いクエリは意味的に関連していると仮定して，クエリの出現頻度の時間的相関を求めるには単純の相関係数が使われている．本研究はトピックが時系列データに意味的に関連しているかどうかを判断するためにトピックの出現頻度の時間変動を利用する点で同じであるが，ニュースが時系列データへの影響の有無を判断するため，クエリの間の類似性を判断するより複雑である．

3. 関連トピックの発見

本章は入力された時系列データに意味的に関連す

るトピックを検出し，ランキングする手法について述べる．

まず，集めたニュース記事をトピックごとに分類する．そして，入力時系列データとの類似度のみで，意味的に関連するトピック候補を抽出する．最後に，類似度と時間的相関の二つの尺度より，トピック候補をランキングする．

3.1. トピックの分類

本研究は入力時系列データに意味的に関連するトピックを提示することによって，関連ニュース記事を提示するため，最初にニュース記事がトピックに割り当てられていることを前提としている．ニュース記事をトピックごとに分類する手法の多くはクラスタリング法である．クラスタリング法は単語の出現状況が似た文章を同じクラスタ（トピック）に分類する．

また，Web ニュースエンジンには注目された話題をトピックや特集にし，その話題に関するニュース記事をまとめて提示するものもある．アサヒコムは最近注目された話題２０個ぐらい「特集」として提示している．Yahoo！ニュースは，ほぼすべてのニュースを「トピックス」に分けられ，全部で 1000 個近くのトピックスがある．ニュース記事のカテゴリ分類に関する方法は本研究の目的ではないため，Yahoo!ニューストピックスで分けられたトピックを使い，分析と説明を行う．

3.2. トピック候補の決定

ニュース記事をトピックごとに分けたら，入力時系列データと関連があるトピックの候補を決める．その方法として，時系列データとそれぞれのトピックの類似度を求め，類似度がある閾値を超えたトピックを候補とする．

この類似度は文書検索で一般的に使用されるベクトル空間モデルに基づいて定義する．文書検索でクエリに関する文書を探すのに対して，ここでは，入力された時系列データの特徴づけキーワード（名前）をクエリに，それに内容的に関連するトピックを検索する．

- **トピックのベクトル表現（*Nf*/*iTpf* 法）**

トピックをベクトルに表現するために，文書のベクトル表現の *tf*/*idf* 法をもとに，*Nf*/*iTpf* 法を定義する．「文書」に対応するのは「トピック」である．トピックが何に関するかを決めるのはその中のニュース記事が何に関するかである．また，ある記事のタイトルに単語 t_j が出現した時，そのニュース記事は単語に関係するものと定義する．したがって，トピックを単語のベクトルで表現し，トピック *Ti* の単語 t_j の *Nf*（News frequency）値はトピック *Ti* におけるタイトルに単語 t_j が出現するニュース記事の総数 *freq*(*i*, *j*) をトピック *Ti* の総記事数で正規化したものである．

$$Nf_{ij} = \frac{\log(\ freq\ (i,j)+1)}{\log(\text{トピック}\ \ i\text{中の総記事数})}$$

単語 t_j の *Tpf* (Topic frequency)値はタイトルに単語 t_j を含むニュース記事を持つトピック数で，単語 t_j の *iTpf* 値は *Tpf* の逆をトピック総数 *N* によって正規化したものである．

$$iTpf_j = \log\frac{N}{Tpf_j}$$

トピック *Ti* の単語 t_j の重み $w_{ij} = Nf_{ij} \times iTpf_j$

- **時系列データ名のベクトル表現**

ユーザによって入力された時系列データの名前中の固有名詞と一般名詞を抽出する．固有名詞に２，一般名詞に１，名前に含まない単語に０というようなベクトルで時系列データ *TS* を表現する．

- **類似度**

トピックと時系列データの類似度 *sim* はそれぞれを表すベクトルの内積で定義する．*Wi* をトピック *Ti* のベクトルで，*Wq* を時系列データのベクトルとしたら，類似度を以下の式で表す．

$$sim(TS,Ti) = w_{q1}w_{i1} + ... + w_{qn}w_{in}$$

時系列データの名前に含まない単語の Q での重みはすべて０であるから，時系列データとトピックの類似度＝2(時系列データの名前にあるすべての固有名詞の *Ti* での重みの和)＋(時系列データの名前にあるすべての一般名詞の *Ti* での重みの和)．

類似度が閾値 *th* を越えたトピックを候補とする．

3.3. トピック候補のランキング

各トピックに入力時系列データとの関連性を表す評価値を与える．評価値はトピックと時系列データの類似度と時間的相関の両方をかけたものにする．類似度は 3.2 で定義したものである．時間的相関によって，トピックが時系列データ全体にどれだけ影響があるかを示そうとしている．基本的な考え方はあるトピックが多く出現するたびに時系列データは大きく変動したら，そのトピックは時系列データに大きく影響する．

したがって，時間的相関を求めるのに，まず各トピックの出現頻度の時系列データを作り，そして，時系列データの変動を表す時系列を求め，最後に二つの時系列データの相関係数を計算する．

トピックの出現頻度の時系列の作成

トピック中のニュース記事を報道する日付ごとにまとめ，その日付の累積ニュース数を数え，各トピックの出現頻度の時系列データを作成する．

3.3.1 時系列データの移動ボラティリティ

ボラティリティとは，変動率の大きさを示す言葉である．時系列データの各時点でのボラティリティを計

算し時間順に並べてできた時系列を移動ボラティリティといい，時系列データの各時点の変動の大きさを示す．例えば，図 1 は時系列データとその移動ボラティリティをグラフにプロットしたものの例である．本研究でボラティリティを計算するのに標準偏差を使う．

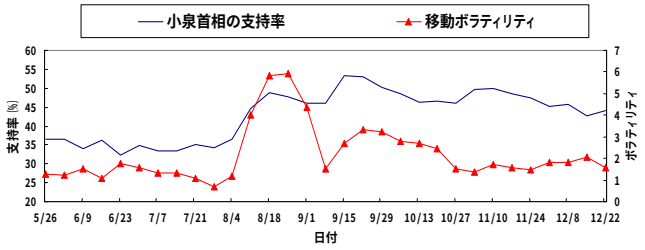

図 1　時系列データと移動ボラティリティの例
（計算期間 $w=5$）

まず，移動ボラティリティの計算期間 w を決める．各日付でのボラティリティはその日付を含み，過去 w 個の時点での値の標準偏差を計算したものである．

3.3.2 時間的相関の計算

トピックの出現頻度の時系列と入力時系列データの移動ボラティリティの相関を計算するため，まず二つの時系列データの時間範囲と時間ステップを同じにする必要がある．

I.　時間範囲と時間ステップの調整

入力時系列データの移動ボラティリティの時間範囲と時間ステップを基準にし，トピックの出現頻度の時系列を修正する．方法として，移動ボラティリティの日付を出現頻度の時系列の日付にし，各日付に対応するトピックの出現頻度はその日付を含み，過去 w（移動ボラティリティの計算期間と同じ値）個目の時点までのトピックの一日あたりの出現頻度の平均値にする．つまり，入力時系列データのある時間帯の変動率（ボラティリティ）にその時間帯のトピックの一日あたりの出現頻度の平均値に対応させる．

トピックはある時間帯にニュース記事が一件もなかったところ，あるいはトピックの出現頻度の時系列の範囲を超えた部分は 0 で埋める．

II.　相関の計算

トピックの出現頻度の時系列と入力時系列データの移動ボラティリティの時間範囲と時間ステップを合わせたら，この二つの時系列データをそれぞれ時間順に並べてできたベクトルの相関係数を求める．

トピックの出現頻度から求めたベクトルを X，入力時系列データの移動ボラティリティからできたベクトルを Y とし，X と Y の相関係数 r を以下の式のように定義する．

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(Xi - \overline{X})(Yi - \overline{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Xi - \overline{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Yi - \overline{Y})^2}}$$

r は -1 から 1 の間の実数である．r は 1 に近いほど，ベクトル X と Y の相関が強い．このとき，トピックが多く出現するとき，時系列データも大きく変動することも多く，トピックが出現しないとき，時系列データもあまり変動しないことが多い．つまりこのトピックの出現は時系列データを変動させる．r は 0 に近づくと，X と Y は無関係になる．このとき，トピックの出現と時系列データの変動とは関係ない．r は -1 に近いほど，X と Y は逆方向で相関が強い．このとき，トピックが多く出現するとき，時系列データはあまり変動せず，トピックが出現しないと，時系列データは大きく変動するようになる．実はこれもトピックが時系列データを変動させないため，関係がないと考えられる．

したがって，トピック Ti と時系列データの時間的相関値 $Correl(TS,Ti)$ は r で定義する．$Correl$ が大きいトピックと時系列データの例を図 2 で示す．

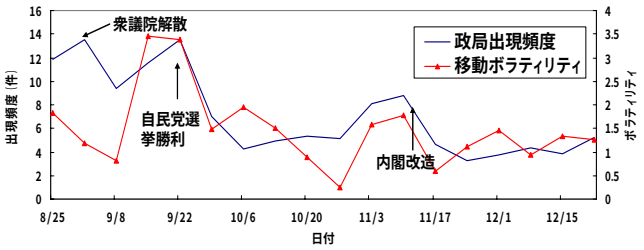

図 2　「小泉首相の支持率」の移動ボラティリティと「政局」の出現頻度

ここで注意すべきなのはトピックと入力時系列データの時間的相関を計算するには入力時系列データそのものではなく，入力時系列データの変動を表す移動ボラティリティである．入力時系列データを用いてトピックの出現頻度と計算した相関係数が大きいことは，トピックの出現頻度と入力時系列データが同時に増やしたら，減らしたりすることを示す．しかしながら，あるトピックがある時間帯で時系列データに影響することを簡単に描くと，図 3 のようであって，同時に上昇するが同時に低下しない．したがって，入力時系列データの変わりにその変動との相関係数を使用する．

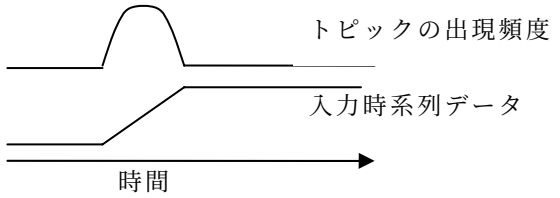

図 3　トピックの出現が時系列データへ影響すること

トピックの評価値

トピック Ti の評価値 Vi は以下の式で計算する．

$$Vi = sim(TS, Ti) \times correl(TS, Ti)$$

各トピック候補に評価値を与えられたら，その評価値によってトピック候補をランキングして提示する．

4. トピックの時間帯別影響の算出

3 章でトピックは入力時系列データ全体の変動に影響するかどうか，またどれだけ影響するかを計算する手法を示した．しかし，どの時間帯で変動させたか，またどうやって変動させたか，上昇させたかあるいは低下させたかはまだ示されていない．そして，注意するのはトピックが異なる時間帯で入力時系列データへ異なる影響を及ぼす可能性があるので，それぞれの時間帯を別々に処理する必要があると考えられる．

本研究の基本的な考えは図 3 で示したように，トピックが急に出現することは時系列データに影響を与える必要条件である．したがって，まずトピックの出現頻度の急に上昇したところ（バースト時間帯）を抽出する．そして，各時間帯で影響を与えているか，またどうのように影響するかを決める．

4.1. 出現頻度のバースト時間帯の抽出

ある時系列データのバースト時間帯は以下の手順によって抽出される．

I. 時系列データの移動平均の計算

時系列データの移動平均(Moving Average)とは時系列データの動向を把握しやすくするため，時系列データの変動をなめらかにしたものである．各時点での移動平均値はこの時点を中心に前後何（v）日分の値の平均値で計算する．

II. 閾値の設定

閾値 *cutoff* を移動平均の平均 *average* と 0.5 倍の標準偏差 *std* の和とする．

$$cutoff = average(MA_v) + 0.5 \times std(MA_v)$$

III. バースト時点の抽出

トピックの出現頻度の各時点の移動平均値について，移動平均値が閾値を超えた時点をバースト時点とする．

図 4 はトピックの出現頻度のバースト時間帯の抽出の例を示す．移動平均の計算期間 $v = 7$ 日としている．

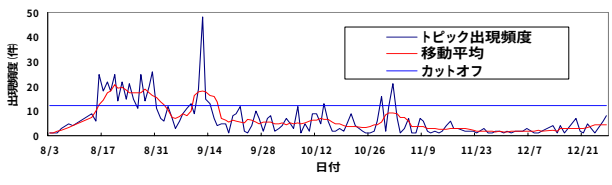

図 4 「郵政民営化」の出現頻度とバースト時間帯(v=7)

4.2. 各バースト時間帯での影響の算出

トピックの出現頻度のバースト時間帯が抽出されたら，各バースト時間帯は時系列データに影響したか，またどのように影響するかは以下の手順で求める．

I. 入力時系列データの移動平均を求める

計算期間は出現頻度の時系列の移動平均を計算するときと同じ計算期間とする．

II. 閾値 θ を決める

入力時系列データの移動平均の標準偏差を閾値 θ とする

III. 各バースト時間帯の影響の求め方

各バースト時間帯に対して，そのバースト時間帯の開始時点の一個前の時点および終了時点に対応する入力時系列データの移動平均の値 s と t を求める．入力時系列データの移動平均の時系列にそれらの時点に対応する日付がない場合，その時点を超えた一番近い日付の値にする． $x = s - t$ とる．

- x は正でかつその絶対値は θ 以上の場合，このバースト時間帯でトピックは時系列データを上昇させた．
- x は負でかつその絶対値は θ 以上の場合，このバースト時間帯でトピックは時系列データを低下させた．
- x の絶対値は θ 未満の場合，このバースト時間帯でトピックが時系列データへ影響を与えない．

5. 実装と実験

3 章と 4 章で述べた手法に基づき，プロトタイプを実装し，検証実験を行った．本章は実装と実験について述べる．

5.1. システムのプロトタイプ

本システムのプロトタイプを図 5 で示す

図 5 に示したように本システムは主に四つのコンポーネントを含む．

I. **入力エリア**：時系列データの名前（ファイル名）をユーザが入力する．

II. **関連トピックエリア**：入力された時系列データと関連があると判断されたトピックをランクして表示する．

III. **トピック分析エリア**：二つの部分から成り立つ．

A) 影響の詳細表示エリア：選択されたトピックの時系列データに影響を与える時間帯とその時間帯でどのように時系列データを変動させたかを表示する．

B) グラフ表示エリア：時系列データやその変動と関連トピックエリアで選択されたトピックの出現頻度の時間的変化を同時にグラフで表示する．曲線上の点をクリックすることによって，対応する日付のニュースを表示可

能である．

IV.　ニュース記事表示エリア

A)　記事タイトル：グラフ表示エリアで選択された日の記事タイトルリストを表示する．

B)　記事本文：記事タイトルで選択された記事の Web ページを表示する．

図 5　システムのプロトタイプ

5.2. 実験

本研究は上で紹介したシステムを用い，実証実験を行った．

5.2.1. 実験データ

実験に使用したのは，「Yahoo!トピックス」から集めた 8 月 5 日から 12 月 28 日までの四ヶ月弱のニュース記事である．トピックは Yahoo!トピックスで分類されたトピックを利用している．「Yahoo!トピックス」はいくつかのカテゴリに分けられた 1000 個近くのトピックスを持ち，5000 個以上のニュースソースから集めたニュース記事をそれらのトピックに割り当てる．そして，新しく注目を浴びる話題が現れるたびに，新しいトピックが作られ，さらに「Yahoo！トピックス」に取り上げられているトピックは，いずれも長時間にわたり話題になっているニュースのカテゴリになっており，本研究で扱うトピックの定義に一致している．

5.2.2. 実験パラメータ

実験時に使われた，提案手法のパラメータは以下のように決めている：

- トピック候補を抽出するための閾値 *th*

3.2 節で時系列データとトピックの類似度によりトピック候補を決める．このステップは入力時系列データとまったく関係のないトピックを取り除き，次のランキングステップで関係がないのにたまに時系列データの変動波形に相関しているようなトピックのランキングに影響を与えないようにする．さらに，閾値をあまり大きくすると文書の類似度は小さいが実は時系列データに影響を与えているようなトピックを見落とさないように設定する必要がある．実験の結果，$th=1$ と決めている．

- 移動ボラティリティの計算時間 *w*

3.3 節で相関係数を計算するとき，移動ボラティリティとそれに対応するトピックの出現頻度の計算期間は実験結果に大きく影響する．長い計算期間は入力時系列データの変動の大きな傾向への影響を重視し，短い計算期間は短時間の変動への影響を重視する．本研究は w を 3 週間程度に設定する．例えば，週ごとの入力時系列データの場合，$w=3$ と設定する．

- 移動平均の計算時間 *v*

4 章で移動平均の計算期間はバースト時間帯の抽出とバースト時間帯の効果の分析に影響を与える．本実験は移動ボラティリティと同様に *v* を 3 週間程度にする．

5.2.3. 実験結果

入力時系列データの二つの例の実験結果を紹介する．

- **小泉首相の支持率**

本実験は小泉首相の 8 月 11 日から 12 月 22 日までの週ごとの支持率（ネット調査 iMi 声活エンジン[13]より）を利用する．この期間中に郵政法案，衆議院解散，選挙の勝利などで支持率は大きく影響が受けた．

Step 2 で 1000 個のトピックから 128 個のトピック候補を得た．

表 1 は評価値，類似度，時間的相関の尺度でそれぞれ上位 10 個のトピックを表す．評価値や時間的相関が 0 以下のトピックは入力時系列データに関しているが影響を与えてないトピックである

表 1 の中で，時系列データの類似度で上位を示したトピックはだいたい小泉首相が取った行動，あるいは小泉首相が大きく関わったことであり，支持率の変動には関係ない順を示す．一方，時間的相関でこの期間（8 月 11 日～12 月 22 日）の支持率の変動と相関があるトピックが上位に上がる．たまに波形の間相関があるが，実は関連が薄いトピック（例，「鉄道事故」）も上位にあがるが，これは類似度をかけることによって小さい評価値を示す．評価値は類似度と時間的相関を合わせた結果である．ユーザはこの結果に基づいて，上位トピックの出現頻度のピークになった時間帯のニュース記事を読むことにより，どんなことが小泉首相のこの期間の支持率に大きく関連しているかがわかる．

例に「郵政事業民営化」で抽出したバースト時間帯を図 6 に示す．この時間帯（8 月 21 日～9 月 22 日）で

記事の平均出現頻度は 11 件で,支持率を上昇させる効果を持つと考えられる.

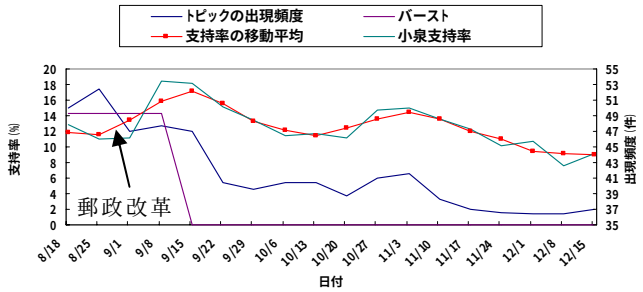

図 6　「小泉首相の支持率」の移動平均と「郵政民営化」の出現頻度のバースト時間帯

- **ブッシュ大統領の支持率**

本実験は米ブッシュ大統領の 8 月 22 日から 12 月 19 日までの 4 日ごとの支持率（PollingReport．com[14]より）を利用する．この期間中の前半でハリケーン・カトリーナ対応への不満，CIA 工作員名の氏名リーク事件により政府への不信感から，支持率は就任以来の最低にもなった．そして，12 月の中旬に入り，イラク選挙成功の影響を受けてやっと上昇へ向かっていた．

入力時系列データは 4 日ごとのデータのため，$w=5$ とする．Step2 で 1000 個のトピックから 106 個のトピック候補を抽出した．

表 2 は評価値，類似度，時間的相関でそれぞれ上位 10 個に現れるトピックを示す

表 1　「小泉首相の支持率」の関連トピック

評価値により				類似度により				時間的相関により			
トピック名	評価値	類似度	相関	トピック名	評価値	類似度	相関	トピック名	評価値	類似度	相関
国内政局	2.70	4.71	0.57	ポスト小泉	-1.43	5.92	0.24	国連安保理改革	1.16	1.58	0.73
郵政事業民営化	2.03	4.59	0.44	靖国神社参拝問題	-2.34	5.42	-0.43	国内政局	2.70	4.71	0.57
日朝国交正常化交渉	1.55	3.62	0.43	小泉純一郎内閣	0.50	5.42	0.09	郵政事業民営化	2.03	4.59	0.44
国連安保理改革	1.16	1.58	0.73	国内政局	2.70	4.71	0.57	日朝国交正常化交渉	1.55	3.62	0.43
選挙	1.09	2.92	0.37	郵政事業民営化	2.03	4.59	0.44	核開発問題	0.99	2.47	0.40
核開発問題	0.99	2.44	0.40	日中関係	-0.26	3.81	-0.07	風水害	0.70	1.82	0.38
風水害	0.70	1.82	0.38	特殊法人改革	-0.18	3.96	-0.05	選挙	1.09	2.92	0.37
日本国憲法	0.66	3.28	0.20	消費税引き上げ問題	-0.26	3.81	-0.07	邦人の事件事故	0.35	1.04	0.33
小泉純一郎内閣	0.50	5.42	0.09	三位一体改革	-1.30	3.73	-0.35	国連(UN)	0.25	1.10	0.22
世論調査	0.41	2.13	0.19	政府系金融機関改革	-1.31	3.66	-0.36	鉄道事故	0.30	1.43	0.21

表 2　「ブッシュ大統領の支持率」の関連トピック

評価値より				類似度より				時間的相関より			
トピック名	評価値	類似度	相関	トピック名	評価値	類似度	相関	トピック名	評価値	類似度	相関
核開発問題	2.82	4.56	0.62	ブッシュ政権	0.28	5.95	0.05	中東情勢	1.75	2.77	0.63
ハリケーン	2.24	3.81	0.59	対テロ戦争	-0.34	5.53	-0.06	核開発問題	2.82	4.56	0.62
国連(UN)	1.86	3.15	0.59	イラク	-1.94	5.12	-0.38	国連	1.86	3.15	0.59
北朝鮮	1.86	3.67	0.51	イラク復興	-1.88	4.64	-0.40	核兵器	1.70	2.88	0.59
中東情勢	1.75	2.77	0.63	核開発問題	2.82	4.56	0.62	ハリケーン	2.24	3.81	0.59
核兵器	1.70	2.88	0.59	イラク戦争	-0.98	4.55	-0.22	韓国経済	0.91	1.55	0.59
テロリズム	1.20	3.78	0.32	米軍動向	0.60	4.17	0.15	世論調査	0.66	1.15	0.58
CIA 工作員名漏えい疑惑	1.14	3.36	0.34	ハリケーン	2.24	3.81	0.59	北朝鮮住民亡命問題	1.01	1.80	0.56
南北朝鮮関係	1.07	2.00	0.53	テロリズム	1.20	3.78	0.32	南北朝鮮関係	1.07	2.00	0.53
北朝鮮住民亡命問題	1.01	1.80	0.56	北朝鮮	1.86	3.67	0.51	北朝鮮	1.86	3.67	0.51

表 2 からは表 1 と同じことが見られる．ただ，「北朝鮮核開発問題」はこの期間中の支持率にそれほど影響がないと思われるが，1 位を示している．実際のグラフを見て，ブッシュ大統領の支持率は二回大きく変動する（下がる）ところで，二回とも，北朝鮮核開発問題も多く報道された（図 7）．このようなことが起きたのはデータの期間が短いのが一つの原因である．データの期間が長ければ，ほかの支持率が大きく変動す

るところでこのトピックが多くならなかったら，時間的相関が小さくなる．もう一つの原因に，もしかして「北朝鮮核開発問題」はブッシュ大統領の支持率にそれほど影響がないと思われるが，実は大きく影響を及ぼしている可能性がある．こういうトピックが見つかるのは，本研究の特徴の一つである．

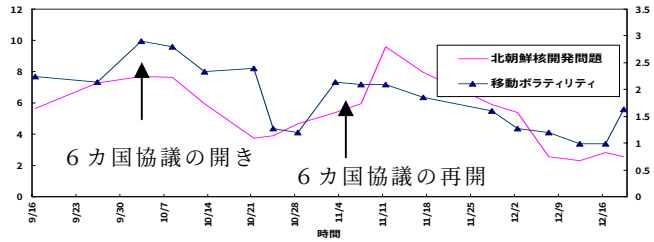

図 7　「ブッシュ大統領の支持率」の移動ボラティリティと「北朝鮮核開発問題」の出現頻度

6. まとめと今後の課題

本稿では文書の類似度と時間的相関の二つの尺度からニューストピックと入力時系列データの関連の有無を判断する手法を提案した．さらに，特定のトピックは入力された時系列データにどの時間帯でどのように変動させるかを求める手法を提案した．

今後の課題として，トピックと入力された時系列データの間に時間的ずれがある場合の処理が考えられる．例えば，「阪神電鉄」の株価の時系列データと「村上ファンド」というトピックは関連が高いと思われるが，本手法を使ったら時間的相関が 0 で影響がないと判断される．原因として，図 8 に示されるように村上ファンドの阪神電鉄株の買い増しに関する記事が多くなったのは買い増しの後であることが考えられる．今後はこのような状況にも対応する手法を開発していく．また，同じトピックで，違う影響を及ぼす時間帯にあるニュース記事の違いを探す機能を拡張していく予定である．

さらに，本論文はユーザから受け取った時系列データの意味的に関連する記事を提示することができたが，これをさらに拡張してニュース記事の関連記事を提示することを考えている．ある記事の過去の内容を時系列で表すことができたら，それを入力として，本提案手法により，その記事の意味的関連記事を発見することを今後の課題として検討する予定である．

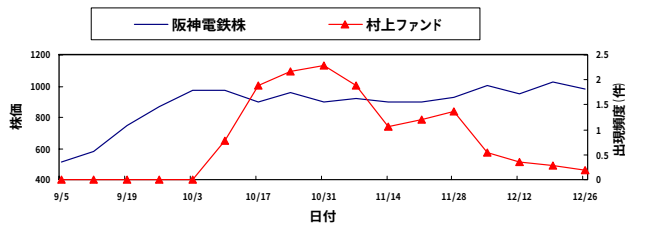

図 8　「阪神電鉄株」と「村上ファンド」の出現頻度

謝辞

本研究の一部は，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形 成」、文部科学省科学技術振興費プロジェクト「異メディア・アーカイブの横断的検索・統合ソフトウェア開発」（代表：田中克己），および，平成 17 年度科研費特定領域研究(2)「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」（課題番号：16016247，代表：田中克己）によるものです．ここに記して謝意を表すものとします．

文　献

[1] http://www.e-marketing-news.co.uk/Aug04-Greg.html

[2] マイボイスコム
http://www.myvoice.co.jp/biz/surveys/8106/index.html

[3] Yahoo!ニュース　http://news.yahoo.co.jp.

[4] J. Allan,J. Carbonell, G. Doddington, J. Yamron, and Y. Yang: “Topic Detection and Tracking Pilot Study Final Report,” Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pp.194-218, Lansdowne, Virginia, USA, 1998.

[5] J. Allan, R. Papka, V. Lavranko, “On-line New Event Detection and Tracking”, Proceedings of 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 37-98, Melbourne, Australia, 1998.

[6] A. Nadamoto, K. Tanaka, “Time-based Contextualized-News Browser (T-CNB)”, Proceedings of the 13th International World Wide Web Conference, pp. 458-459, New York, USA, May, 2004.

[7] J. Allan, V. Khandelwal, and R. Gupta, “Temporal Summaries of News Topics”, Proceedings of the 24th annual international ACM SIGIR conference on Research and development of information retrieval, pp. 10-18, New Orleans, Louisiana, USA, 2001

[8] F. Fornari. C. Monticelli, M. Pericolli and M. Tivegna, “The Impact of News on the Exchange Rate of the Lira and Long-Term Interest Rates,” Economic Modeling, Elsevier, Vol. 19, No. 4 pp. 611-639, 2002.

[9] K. Kauffmanand A. Weerapana, “The Impact of AIDS – Related News on Exchange Rates in South Africa,” Working paper, Wellesley College, 2005.

[10] S. DellaVigna, E. Kaplan, “The fox news effect: Media bias and voting”, Working Paper, UC Berkeley, 2005.

[11] S. Chien, N. Immorlica, “Semantic Similarity Between Search Engine Queries Using Temporal Correlation”, Proceedings of the 14th International World Wide Web Conference, pp. 2-11, Chiba, Japan, May, 2005.

[12] R. Papka and J. Allan, “On-line new event detection using single-pass clustering”, Technical Report UMASS Computer Science Technical Report 98-21, Department of Computer Science, University of Massachusetts, 1998.

[13] iMi 声活エンジン
http://www.imi.ne.jp/abc/cgi/ise_genre.cgi

[14] PollingReport.com
http://www.pollingreport.com/BushJob.htm