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

Discovery of Semantically Related Topics for Given Time Series Data

張 一萌♥ 何 書勉♦ 小山 聡
田島 敬史 田中 克己♠

Yimeng ZHANG Shumian HE
Satoshi OYAMA Keishi TAJIMA
Katsumi TANAKA

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．これによって入力された時系列データの変動に影響を及ぼすような関連トピックを検出する．さらに，特定のトピックに関する記事の出現頻度の時間的特徴により，そのトピックがどの時間帯でどうのように入力された時系列データを変動させたかを分析する．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．

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 topics which have some impact on the time series data. Moreover, 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.

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．しかし，現在のニュースサイトは出来事が報道されているだけで，テレビニュースのような専門家による出来事の分析は稀である．ニュースサイトの場合は，専門家による分析の代わりに，インターネット上の大量のニュース記事に情報技術を適用することで，自動的に社会的性質の分析ができるのではないかと考えられる．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ニュース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，郵政民営化など）のことである．

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，そのトピックは時系列データの変動に影響を与えていると仮定する．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．

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[1]がある．本研究ではトピック検出より検出し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[2][3]．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い，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[4][5]．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られるクエリ間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[6]．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. 提案手法

本研究はユーザから時系列データとその特徴づけるキーワード（時系列データの名前）を受け取り，トピックと時系列データの類似度と時間的相関の二つの尺度より，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．さらに，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どうやって変動させたか，つまり上昇させたかある

♥ 学生会員 京都大学大学院情報学研究科修士課程
zhangym@dl.kuis.kyoto-u.ac.jp
♦ 学生会員 京都大学大学院情報学研究科博士課程
shumian@dl.kuis.kyoto-u.ac.jp
♠ 正会員 京都大学大学院情報学研究科
{oyama, tajima, tanaka}@dl.kuis.kyoto-u.ac.jp

いは低下させたかを判断する．本章はそれらを実現する提案手法を紹介する．

3.1 関連トピックの発見

まず，集めたニュース記事をトピックごとに分類する．これは本研究の目的ではないため，Yahoo!ニューストピックス[1]で分けられたトピックを使い，分析と説明を行う．Yahoo！ニュースはほぼすべてのニュースを 1000 個近くのトピックスに割り当てる．

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．

3.1.1 トピック候補の決定

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．文書検索ではクエリに関連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．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**

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すると，トピックと時系列データの類似度 *sim* はそれらの内積で定義し，以下の式で表す．

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

類似度が閾値 *th* を越えたトピックを候補とする．ただし，閾値をあまり大きくすると，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す可能性があるため，閾値をそれほど大きくないように設定する．

3.1.2 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 3.1.1 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，まず入力された時系列データの変動を表す時系列（時系列データの移動ボラティリティ）を求める．移動ボラティリティは時系列データの各時点での変動率の大きさ（ボラティリティ）を時間順に並べてできた時系列データである．ここで，入力された時系列データの各日付でのボラティリティはその日付を含む過去 *w* 個の時点での値の標準偏差とする．*w* は先に決めた移動ボラティリティの計算期間である．

次に，トピックに関する記事の出現頻度の時系列データを求める．方法として，入力された時系列データの移動ボラティリティの日付を出現頻度の日付にし，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．

そして，この二つの時系列データを時間順に並べたベクトルの相関係数を求める．トピックの出現頻度の時系列データ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．

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．

3.2 トピックの時間帯別影響の算出

節 3.1 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．

[1] http://dailynews.yahoo.co.jp/fc/

本研究の基本的な考えは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．

出現頻度のバースト時間帯の抽出

まず，トピックに関する記事の出現頻度の時系列データの移動平均を求める．各時点での移動平均値はその時点を中心に前後何（v）日分の値の平均値で計算する．

次に，閾値を決める．閾値 $cutoff$ を以下のように設定する．

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

α は実験の結果，0.5 は最適である．

そして，トピックに関する記事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．

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．まず，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．

そして，入力時系列データの移動平均の標準偏差を求め，閾値 θ とする．

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．

4. 実験と考察

4.1 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 2005 年 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る．

4.2 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[7]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．

表 1 は評価値，類似度，時間的相関の尺度でそれぞれ上位 8 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も上位に上がるものもあるが，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．

- **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[8]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．

表 2 は評価値，類似度，時間的相関でそれぞれ上位 8 個に現れるトピックを示す．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 1）．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．

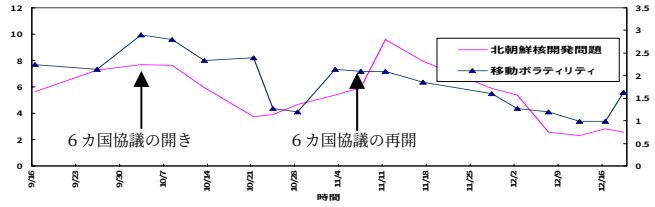

図 1 「ブッシュ大統領の支持率」の移動ボラティリティと「北朝鮮核開発問題」の出現頻度

Fig.1 Moving volatility of “President Bush’s approval ratings” and frequency of “North Korea nuclear program”.

5. 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．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

表１　「小泉首相の支持率」の関連トピック
Table1 Related topics of “Premier Koizumi’s approval ratings”.

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	類似	相関	トピック名	評価	類似	相関	トピック名	評価	類似	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	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	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	-0.26	3.81	-0.07	邦人の事件事故	0.35	1.04	0.33

表２　「ブッシュ大統領の支持率」の関連トピック
Table2 Related topics of “President Bush’s approval ratings”.

評価値より				類似度より				時間的相関より			
トピック名	評価	類似度	相関	トピック名	評価	類似	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	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

することを今後の課題として検討する予定である．

[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 18 年度科研費特定領域研究「情報爆発に対応するコンテンツ融合と操作環境融合に関する研究」（課題番号：18049041，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] 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.

[2] 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.

[3] 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

[4] 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.

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

[6] 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.

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

[8] PollingReport.com http://www.pollingreport.com/

張 一萌 Yimeng ZHANG
2006 年京都大学工学部情報学科卒業．同年，同大学院情報学研究科修士課程入学，現在に至る．日本データベース学会学生会員．

何 書勉 Shumian HE
2004 年京都大学大学院情報学研究科修士課程修了．同年，同大学院情報学研究科博士課程入学，現在に至る．ユビキタスコンピューティングに関する研究に従事．情報処理学会，日本データベース学会学生会員．

田島 敬史 Keishi TAJIMA
1991 年東京大学理学部情報科学科卒業．1994～1996 年京都大学数理解析研究所研究生．1996 年東京大学理学系研究科情報科学専攻博士課程修了．博士(理学)．神戸大学助手，ペンシルバニア大学客員助手，北陸先端科学技術大学院大学助教授を経て，2005 年より京都大学情報学研究科助教授．主にデータベースシステム，Web 検索の研究に従事．ACM，情報処理学会，ソフトウェア科学会，日本データベース学会，各会員．

小山 聡 Satoshi OYAMA
京都大学大学院情報学研究科社会情報学専攻助手．2002 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主に機械学習，データマイニング，情報検索の研究に従事．電子情報通信学会，情報処理学会，人工知能学会，日本データベース学会，IEEE，ACM，AAAI 各会員．

田中 克己 Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．京大工博．主にデータベース，Web 情報検索，マルチメディアコンテ ンツ処理の研究に従事．IEEE Computer Society, ACM，人工知能学会，日本ソ フトウェア科学会，情報処理学会，日本データベース学会等各会員．