DEIM Forum 2011 E2-4

ニュース情報伝播の時空間的局所性の分析手法

山口　彰太† 　是津　耕司†† 　木俵　豊†† 　田中　克己†

† 京都大学大学院情報学研究科社会情報学専攻　〒 606–8501 京都府京都市左京区吉田本町
†† 情報通信研究機構知識創成コミュニケーション研究センター　〒 619–0289 京都府相楽群精華町光台 3-5
E-mail: †{syamaguchi,tanaka}@dl.kuis.kyoto-u.ac.jp, ††{zettsu,kidawara}@nict.go.jp

あらまし　インターネットの普及に伴い，実世界の様々な出来事に関するニュース情報を世界中の誰もが同時に気軽に手に入れられるようになったが，大量に発信される情報の中から情報を取捨選択するようになったため，ある時間や地域において伝播している情報に偏りが発生している問題がある．本稿では，ある話題に関する文書から内容の主題となるテーマを抽出し，そのテーマの情報伝播に時空間的局所性があるか分析する手法を二つ提案する．一つ目の手法として，抽出したテーマが伝播している地域や時間を抽出することで，時空間的局所性を求める手法である．二つ目の手法として，抽出したテーマの地域や時間ごとにより出現確率を求め，それらの出現確率の値を比較することで，時空間的局所性を求める手法である．この提案手法の有用性を検証するため，全世界的な関心の高さを示したチリの鉱山事故のニュース記事を用いて情報伝播の時間的，地域的な局所性の分析と評価を行う．

キーワード　時空間情報，確率分布，情報伝播の数値化と分類，情報伝播の比較，時空間パターン

1. はじめに

インターネットの普及に伴い，実世界の様々な出来事に関するニュース情報が，Web 上に日々大量に発信されるようになった．これにより，世界中の誰もがニュース情報を同時に手に入れることが出来るようになった．一方，情報が大量に発信されたことにより，ユーザはこれらの情報の全てを把握することが困難になり，多様化された情報の中から情報を取捨選択するようになった．その結果，世界中の誰もが情報を同時に手に入れられるのにも関わらず，日本でしか話題になっていない情報や，2010 年 9 月にしか話題になっていない情報のみを入手するなど地域や時間によって得られる情報の偏りの局所性が存在するようになった．そのため，ある話題に関する情報をまんべんなく知りたいユーザは，情報の網羅性を高めるために，様々な Web サイトを閲覧して，そのサイトにしか存在しない情報や，重複した情報などを取捨選択する必要がある．

本研究は，情報伝播の時空間的局所性が存在するときにユーザが得る情報に局所性が存在すると仮定し，情報伝播の時空間的局所性を分析することで，局所的にしか伝えられていない情報を検出することを目的としている．局所的にしか伝えられていない情報の検出が可能となれば，これら局所的な情報をユーザに提示することで，その地域や時間に伝えられていない情報を相互に補完することが可能となり，ユーザが網羅性高く情報を収集できるようになる．情報伝播の時空間的局所性を分析するためには，ある時刻や場所においてどのような情報が伝播しているのかを検出する必要がある．そのために，世界的関心事を伝えるニュース情報に対し，伝播している情報を言語を超えて扱う必要があり，情報源を偏りなく選ぶ必要がある．また，ニュース情報が指す時間や空間を特定し，それらの情報伝播を調べることで，そのニュース情報がカバーしている時間や空間的な範囲の特定，伝播したニュース情報の差異など情報伝播の網羅性を求める必要がある．

本研究では，情報伝播の時空間的局所性を分析するために，情報伝播の網羅性に焦点を当てる．ある話題に対する情報が特定の時刻や地域において，伝播しているのかを検出するために，その時刻や地域における情報の出現確率を条件付き確率を用いて算出し，確率分布を求める．さらに，求めた確率分布の値を用いることで，特定の時刻や場所間の情報伝播の比較を行い，この中から有意な違いを時空間的局所性として検出する手法を提案する．

本研究では，全世界的な関心を集めたチリの鉱山事故に関するニュース情報を用いることで，チリの鉱山事故に関するテーマの時空間的局所性の分析を行い，分析手法の評価と情報伝播の時空間的局所性の検出を行う．さらに，得られた局所性な情報を，ユーザに推薦するための課題を示す．

本稿の構成は以下の通りである．まず 2. 節では本研究の基本となる概念を述べ，3. 節では本研究で用いる手法や手順を述べる．4. 節では予備実験として，チリの鉱山事故に関するニュース記事に本研究を適用した場合の実験方法とその結果を述べ，5. 節でその考察を述べる．7. 節で今後の研究課題を述べ，6. 節で，ユーザに情報を推薦するためのアプリケーション例とその課題を述べる．8. 節では，時空間情報を用いたパターンマイニングの関連研究の紹介と本研究との相違点を述べ，最後に 9. 節で本稿のまとめを述べる．

2. 基本概念

ニュース情報の時空間的局所性を検証する取り組むべき課題として，以下の課題が挙げられる．

（1） 言語の違いを超えて情報を扱う

（2） 情報源を偏りなく選ぶ

（3） ニュース情報が指す時間・空間を特定する

（4） 情報伝播の網羅性を調べる

本研究では，4 の課題について焦点を当てる．

2.1 定　　義

本研究では，チリの鉱山事故など特定の話題に対する文書を対象とし，チリの鉱山事故の被害状況や，救済方法などのサブトピックの事をテーマ θ として扱う．

本研究では，情報の時空間的局所性を分析するために，時空間情報が記載されている文書のみを対象とする．この時，対象とする文書集合 D を $\mathrm{D} = \{d_1,d_2,\cdots,d_i\}$ として定義する．この集合に加え，各文書が書かれた時刻の集合である T を $\mathrm{T} = \{t_1,t_2,\cdots,t_j\}$ とし，各文書が書かれた場所の集合である L を $\mathrm{L} = \{l_1,l_2,\cdots,l_k\}$ として定義する．さらに，各文書の内容の主題となるテーマの集合 Θ を $\Theta = \{\theta_1,\theta_2,\cdots,\theta_i\}$ として定義する．テーマ集合の各要素 θ は，文書集合 D に出現する単語の集合 $\mathrm{W} = \{w_1,w_2,\cdots,w_n\}$ の一つ以上の要素からなる集合とする．これらの集合が存在するとき，文書集合 D のそれぞれの要素について，その文書が書かれた時刻や場所，その内容の主題となるテーマをラベル付けした集合を文書集合 $\mathrm{C} = \{(d_1,t_1,l_1,\theta_1),(d_2,t_2,l_2,\theta_2),\cdots,(d_i,t_j,l_k,\theta_l)\}$ として定義する．

文書集合 C からあるテーマについて書かれている文書が出現する確率を求める方法を示す．文書 d_1 の内容の主題となるテーマは θ_1 であるから，文書 d_1 は θ_1 に関心があるとみなすことが出来る．このことから，θ_1 への関心の強さを検出するには，θ_1 を主題とする文書の数で示すことが出来る．ここで，時刻 t_1 と時刻 t_2 における θ_1 への関心の強さを比較する場合を考える．時刻 t_1 における文書数 n_{t_1} と t_2 における文書数 n_{t_2} は一般的に異なるため，時刻 t_1 における θ_1 をテーマとする文書数 $n_{t_2\theta_1}$ と時刻 t_2 における θ_1 をテーマとする文書数 $n_{t_2\theta_1}$ を比較する場合には，文書数 n_{t_1} と n_{t_2} を正規化する必要がある．これは場所間の比較やテーマ間の比較でも同様のことが言える．そのため，関心の強さを内容の主題を θ とする文書の数で表すのではなく，対象とする文書中で θ を主題とする文書の出現確率で示すこととする．これにより，人々がある話題に対して，どのテーマにどの程度の割合で関心があるかを示すことが出来る．

このことから，ある話題に関する文書の出現確率を確率分布を用いて以下のように定義することが出来る．

$$P(l,t,\theta) \tag{1}$$

(1) 式から，ある場所 l におけるテーマ θ に対する時刻の違いによる関心の示し方を以下のように求めることが出来る．

$$P(t|\theta,l) \tag{2}$$

(2) 式は，地域 l におけるテーマ θ の時間の違いによる出現確率の変化を示したものである．(2) 式を用いることで，場所 l におけるテーマ θ に対する関心の強さの時間の経過による変化を求めることが出来る．さらに，この確率分布から得られるグラフの面積を計算することで，その場所における関心の高さを求めることが出来る．また，場所 l_1 におけるテーマ θ に対する関心の示し方 $P(t|\theta,l_1)$ と場所 l_2 における関心の示し方 $P(t|\theta,l_2)$ を比較することで，テーマ θ に対する関心の示し方の場所 l_1 と場所 l_2 の違いを求めることが出来る．

物理的な時空間距離 ＼ 情報伝播の内容差異	同じ	異なる
近い	d_{11}	d_{12}
遠い	d_{21}	d_{22}

表 1　時空間情報の分類

同様に，(1) 式から，ある時刻 t において，各地域における各テーマへの関心の示し方は以下のように求めることが出来る．

$$P(l,\theta|t) \tag{3}$$

(3) 式は，時刻 t における各地域における各テーマの出現確率の変化を示したものである．(3) 式を用いることで，時刻 t におけるテーマ θ に対する関心の強さの地域の違いによる変化を求めることが出来る．また，時刻 t_1 におけるテーマ θ に対する関心の示し方 $P(l,\theta|t_1)$ と時刻 t_2 における関心の示し方 $P(l,\theta|t_2)$ を比較することで，各地域が各テーマに示している関心の時刻ごとの違いを求めることが出来る．

2.2 ニュース情報伝播の相関関係

本研究では，ニュース情報の伝播と情報の物理的な影響範囲には相関関係があると仮定し，この相関関係を表 1 に示す．d_{11} では，物理的な時空間距離が近く，伝播している情報内容も同じようなものである．これは，災害ニュースの場合において，被害状況を伝える情報などが付近の地域に即座に伝播していると考えることが出来る．d_{12} では，物理的な時空間距離は近いが，伝播している内容情報が異なるものである．これは二次的な情報伝播が発生したと言える．災害ニュースの場合において，直接的な被害でなく，その災害によって引き起こされる二次災害や社会的な影響に関する情報が伝播していると考えられる．d_{21} では，物理的な時空間は遠いが，伝播している情報内容は同じようなものである．災害ニュースの場合では，災害が発生した時空間から離れていても，被害状況などに興味があるような情報伝播が発生していると言える．そのため，この情報伝播はその話題やテーマに関する関心の広がりの影響範囲を示していると考えられる．d_{22} では，物理的な時空間が遠く，伝播している情報も異なっている．災害ニュースの場合においては，その災害に対する印象や感想など，災害に間接的に関する情報が含まれると考えられる．

このように情報を分類することで，d_{12} のような物理的な時空間は近いのに，二次災害など間接的に関係するテーマに注目している情報を抽出できたり，d_{21} のようなその話題に対する関心の時空間的な広がりを抽出できる．また，d_{12} と d_{21} を比較することで，伝播している情報の差異がどのように発生しているのかを知ることが出来る．それにより，ある話題と関係のあるものではあるが，時空間的に近い情報のみを集めていただ

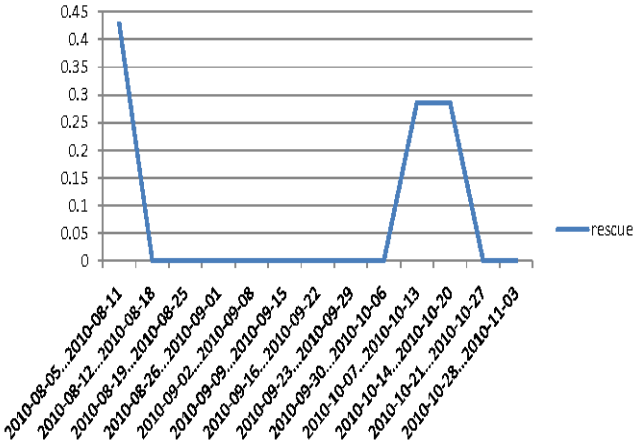

図 1　チリにおける rescue の出現確率

けでは得られない情報を得ることができ，これらの話題のテーマに関する情報の時空間的な影響範囲を知ることが出来る．

2.3　情報伝播の時空間的局所性

本研究で対象とする情報伝播の時空間的局所性について述べる．(2) 式を用いて，各地域における各テーマの出現確率を算出する．ある地域において，そのテーマが出現していれば，そのテーマに関するニュース情報が伝播しているとみなす．これを図 1 を用いて説明する．図 1 は l=chile，θ=rescue とした場合の出現確率の確率分布をグラフ化したものである．この時，chile において，rescue という単語が出現しているのは，t=2010-08-5...2010-08-11 という期間と 2010-10-07...2010-10-20 という期間であることが分かる．このことから，chile において rescue というテーマはこれら二つの期間において情報が伝播していると考えられる．また，同様に，他の地域における rescue というテーマが出現する期間を求めることで，rescue というテーマについて，情報が伝播している地域や期間を求めることが出来る．このようにして，各テーマの伝播している地域や期間を求めることで，そのテーマに関する時空間的局所性を算出する．

3.　提 案 手 法

本節では，情報伝播の時空間的局所性の検出手法について述べる．本稿の情報伝播の時空間的局所性の検出手法として，二種類の手法を提案する．

3.1　ValidTime，ValidSpace を用いる手法

一つ目の提案手法として，あるテーマに関する情報が伝播している地域を ValidSpace，あるテーマに関する情報が伝播している時間を ValidTime と定義することで，あるテーマが有効な時間や地域を比較することで，時空間的局所性を検出する手法である．この手法では次のような仮説を用いる．ある話題に関するニュース情報が生成されている期間の中で，その話題のテーマ θ は次のように分類できる．

（1）　どの期間や地域に対しても常に関心を集めているテーマ

（2）　ある期間やある時間に対して一時的に関心を集めているテーマ

(2) の例として，日本でのみ関心を集めているテーマであったり，2010 年 9 月のみに関心を集めているテーマなどが該当する．また，2010 年 9 月には関心を集めていたが，2010 年 10 月には関心がなくなり，2010 年 11 月に再び関心を集めたテーマなども (2) に該当する．このようにして，情報が伝播した地域や期間を抽出することでグループ分けを行い，それぞれを比較することで，テーマの時空間的局所性を発見する手法である．

まず，情報伝播の ValidTime を求める手法について述べる．ある一つの地域 l_1 内の時刻 t_1 から時刻 t_2 の期間において，同一のテーマ θ_1 に関する情報が伝播している状態を以下のように定義する．

- $P(t_1|l=l_1,\theta=\theta_1) \geqq \alpha$，$P(t_2|l=l_1,\theta=\theta_1) \geqq \alpha$ ($\alpha \geqq 0$)

α は閾値であり，各出現確率が閾値を超えた場合に，地域 l_1 内において，時刻 t_1 から時刻 t_2 の期間において，θ_1 に関するニュース情報が伝播したとみなす．このようにして，地域 l_1 におけるテーマ θ_1 の情報伝播の ValidTime(t_1,t_2) を求める．なお，この ValidTime は (t_1,t_2) の期間だけでなく，(t_3,t_4) など複数の期間をもつものとする．同様にして，地域 l_1 における各テーマの ValidTime を算出する．さらに，各地域における各テーマの ValidTime を算出する．

次に，ValidSpace を求める手法について述べる．あるテーマ θ_1 のニュース情報 (d_1,t_1,l_1,θ_1) が (d_2,t_2,l_2,θ_1) に伝播している状態を以下のように定義する．なお，θ_1 は地域 l_1 の ValidTime として t_1 の時刻を含み，地域 l_2 の ValidTime として t_2 の時刻を含むものとする．

- $P(t_1|l=l_1,\theta=\theta_1) \geqq \beta$，$P(t_2|l=l_2,\theta=\theta_1) \geqq \beta$ ($\beta \geqq 0$)

β は閾値であり，各出現確率が閾値を超えた場合に，地域 l_1 から地域 l_2 に時刻 t_1 から時刻 t_2 の期間において，θ1 に関するニュース情報が伝播したとみなす．このようにして，θ_1 の時刻 t_1 から時刻 t_2 の期間における ValidSpace を求める．なお，この ValidSpace も ValidTime と同様に，(t_1,t_2) の ValidTime において (l_1,l_2) の ValidSpace が存在するだけでなく，$P(t_2|l=l_3,\theta_1) \geqq \beta$ となるような地域 l_3 が存在する場合には，ValidSpace は (l_1,l_2,l_3) など，複数の地域をもてるものとする．また，この閾値を超える地域 l_1 と l_2 の組み合わせが存在しない場合，ValidSpace は単一の地域 (l_1) や (l_2) としてみなす．同様にして，各テーマに対して，ValidSpace とその ValidTime を算出する．

このようにして，あるテーマに関する情報が伝播した地域と時間を求める．あるテーマに関するニュース情報伝播が複数発生している場合，これはそのテーマに関する異なる情報伝播が発生したと考えることが出来る．この時，それぞれの情報伝播の ValidTime と ValidSpace において，あるテーマは時空間的局所性が存在すると判定する．

3.2　出現確率の確率分布を用いる手法

二つ目の手法として，2. 節で述べた (2) 式と (3) 式を用いる手法を提案する．(2) 式から，あるテーマに対する出現確率の差を求めることで，各地域間の出現確率の違いを確率分布を用

いることで表現することが出来る．これらの違いを比較することで，各地域間における空間的局所性の検出を行うことが可能となる．(3) 式から，ある時刻におけるテーマの出現確率の違いを確率分布を用いることで表現することが出来る．また，時刻間で出現確率を比較することにより，時刻間における時間的局所性の検出を行うことが可能になる．

確率分布の比較は以下の手法を用いる．ある話題に対する関心の時間変化による地域的な違いを検証するために比較する地域 l を l_1 と l_2，ある話題 θ を θ_1 とする．この時，それぞれの確率分布は以下のように示される．

$$P(t|\theta = \theta_1, l = l_1) \tag{4}$$

$$P(t|\theta = \theta_1, l = l_2) \tag{5}$$

時間集合 T のそれぞれの要素 t に対して，(4) 式と (5) 式における出現確率を求め，得られた出現確率の差を算出する．これにより，時刻 t における地域 l_1 と地域 l_2 のテーマ θ_1 に対する関心の強さの違いを (4) 式から (5) 式を減算することによって求め，次のように定義することが出来る．

$$\Delta P(t|\theta_1)[l = [l_1, l_2]] \tag{6}$$

(6) 式で示されるような二つの地域間の関心の強さの違いを，地域集合 L の要素の全ての組み合わせ ${}_{|L|}\mathrm{C}_2$ 個の式を導出する．さらに，テーマ集合 Θ の全てのテーマに対しても同様に式を導出する．本稿の目的である局所性を求めるために，ここで得られた ${}_{|L|}\mathrm{C}_2 \times |\Theta|$ 個のそれぞれの式に対して局所性の存在の判定及び評価を行う．(6) 式は地域 l_1 と地域 l_2 のテーマ θ_1 に対する関心の強さの違いの変化を示している．提案手法では，関心の強さの局所性を見るために，(6) 式の値がある時刻 t を起点として急激に変わるような式を局所性が存在する式として扱う．この理由として，(6) 式の値がある時刻 t を起点として急激に変わるような場合，これは比較した地域間でそのテーマに対する関心の示し方が急激に変化したとみなすことが出来るため，比較した地域間で地域による局所性が存在すると判定できるからである．このようにして，得られた ${}_{|L|}\mathrm{C}_2 \times |\Theta|$ 個のそれぞれの式から，任意の時刻 t を起点として急激に出現確率の差の値が変化するような式のみを局所性の評価に用いる．この手法により，情報伝播の時間的局所性を検出する．

また，ある話題に対する関心の地域変化による時間的な違いを検証するために，比較する時刻 t を t_1 と t_2 とする．この時，それぞれの確率分布は以下のように示される．

$$P(l, \theta|t = t_1) \tag{7}$$

$$P(l, \theta|t = t_2) \tag{8}$$

地域集合 L のそれぞれの要素 l に対して，(7) 式と (8) 式における出現確率を求め，得られた出現確率の差を算出する．これにより，地域 l において時刻 t_1 と時刻 t_2 のテーマ θ に対する関心の強さの違いを次のように定義することが出来る．

$$\Delta P(l, \theta)[t = [t_1, t_2]] \tag{9}$$

本稿の目的である局所性を求めるために，ここで得られた (9) 式から関心の強さの変化の局所性の判定及び評価を行う．(9) 式は時刻 t_1 から時刻 t_2 間で，関心の強さが地域的にどのような変化をしたのかということを示している．関心の強さの地域的な変化というのは，あるテーマの情報が地域的にどのように伝播していったのかということと同義である．そのため，(9) 式はあるテーマが時刻 t_1 から時刻 t_2 間で，どのような情報伝播が発生したのかを示している式と言える．また，情報伝播を扱う上で，時刻 t_1 でテーマに関心がある地域の集合をクラスタとみなす．時刻 t_1 のクラスタの状態と時刻 t_2 のクラスタの状態の変化として，内部の関心の強さの変化と，クラスタのサイズや位置の変化の二種類が存在する．クラスタ内部の関心の強さの変化には，以下の三通りの分類が考えられる．

- 時刻 t_1 の関心の強さよりも，時刻 t_2 の関心の強さの方が大きい
- 時刻 t_1 の関心の強さと時刻 t_2 の関心の強さは等しい
- 時刻 t_1 の関心の強さよりも，時刻 t_2 の関心の強さの方が小さい

これらは，時刻 t_1 と時刻 t_2 におけるクラスタに含まれる各地域のテーマの出現確率の差を求めることで分類を行うことが出来る．また，クラスタのサイズや位置などの変化には以下の四通りの分類が考えられる．

- 時刻 t_1 に比べ，時刻 t_2 のときの方がクラスタのサイズが拡大している
- 時刻 t_1 も時刻 t_2 のときも，関心を示しているクラスタの位置もサイズも変わらない
- 時刻 t_1 に比べ，時刻 t_2 のときの方がクラスタのサイズが縮小している
- 時刻 t_1 も時刻 t_2 のときも，関心を示しているクラスタのサイズは同じであるが，位置が変化している

これらの分類を行うことにより，あるテーマに関する情報が空間的にどのように拡散し，関心の強さがどのように伝播したのかを示すことが出来る．この手法により，情報伝播の空間的局所性を検出する．これらの手法により得られた時間的局所性と空間的局所性を地域とテーマごとにみることによって，情報伝播の時空間的局所性を検出する．

4. 予 備 実 験

予備実験として，時空間的局所性の分析を行う手法として提案した 3.1 節の手法を用いて，チリの鉱山事故を話題としたニュース情報を対象として，情報伝播の時空間的局所性を検出する実験を行った．

4.1 実験データ

本稿では，情報伝播の時空間的な局所性を検証するために，世界的な関心を集めた話題を対象として実験を行った．本稿では，2010 年 8 月 5 日に発生したチリの鉱山事故を対象として，世界 11 地域の選定し，各地域に所在するニュースメディアサイトをそれぞれ 1 つづつ指定した．また，今回対象としたニュースメディアサイトは，ニュース記事が英語で記載されているものに限定した．さらにこれらのサイトから 2010 年 8 月 5 日から 2010 年 11 月 3 日までに書かれた記事を対象として，チリの

地域	ニュースメディア	記事数	検索クエリ
Chile	http://www.ilovechile.cl/	17	mining accident
Los Angels	http://www.dailynews.com/	8	chile mining accident
New York	http://www.nydailynews.com/index.html	19	chile mining accident
China	http://www.chinadaily.com.cn/index.html	20	chile mining accident
United Kingdom	http://www.bbc.co.uk/news/	19	chile mining accident
Russia	http://english.ruvr.ru/	17	chile mining
Australia	http://www.theaustralian.com.au/	11	chile mining accident
India	http://www.expressindia.com/	5	chile mining accident
Japan	http://www.japantimes.co.jp/	6	chile mine
Italy	http://www.zenit.org/	25	chile mining accident
SouthAfrica	http://www.mg.co.za/	11	chile mining accident

表 2　取得地域と記事数

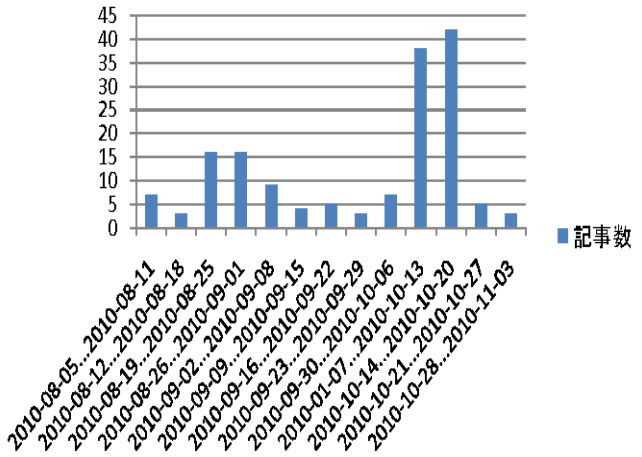

図 2　各期間に生成された記事数

鉱山事故に関するニュース記事のクローリングを行った．今回対象とした地域と，各ニュースメディア，記事数とサイト内検索で用いたクエリを表 2 に示す．本稿では，クローリングから得られた計 158 件のニュース情報に対して，予備実験を行う．

4.2　パラメータの設定

確率分布を求めるために，文書 d から文書が生成された時刻 t と文書が生成された地域 l，その文書の内容の主題であるテーマ θ を求めなければならない．本稿では，取得した文書 d に対し，各ニュースメディアに対応する構造解析を行うことで，文書が生成された時刻 t を取得した．さらに，本稿では，ニュース記事が生成された時刻 t を，一日ごとではなく，一週間ごととして扱った．各期間に生成された記事数を図 2 に示す．文書が生成された地域 l については，各ニュースメディアを予め指定しているため，クローリング元のニュースメディアの所在地を文書が生成された地域 l として取得した．また，文書の主題となるテーマ θ は，以下の仮定を用いて取得した．

- ニュース記事の主題となるテーマ θ は，その文書のタイトル及び本文の一文目に出現する

この仮定により，各文書のタイトル及び本文の一文目に出現する単語からストップワードを取り除いたテーマ集合を各文書ごとに生成した．各文書のテーマはこのテーマ集合に含まれる単語として，テーマ θ を取得した．

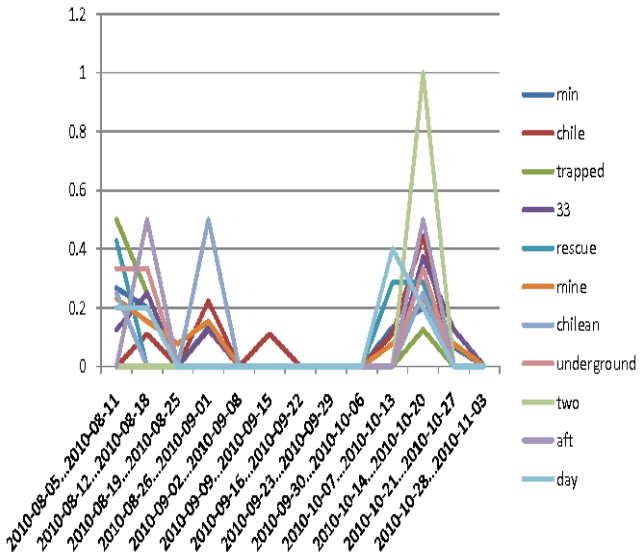

図 3　チリにおける各テーマの出現確率

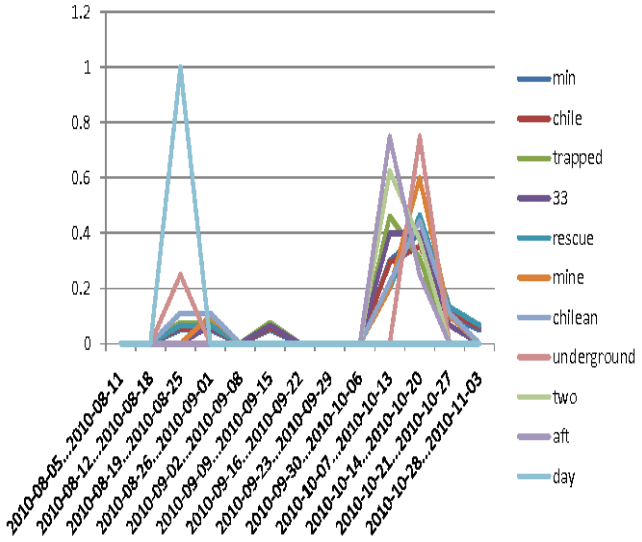

図 4　中国における各テーマの出現確率

4.3　各地域における各テーマの出現確率の確率分布

4.2 節で求めた各値を用いて，(2) 式で各地域と各テーマにおける出現確率を算出した．チリにおける各テーマの出現確率の一部を図 3 に，中国における各テーマの出現確率の一部を図 4 に示す．また，今回の実験では，$\alpha=0$，$\beta=0$ として，その地域や時間にそのテーマが書かれている文書が存在すれば，そのテーマの情報は伝播していることとして扱った．

4.4　得られた時空間的局所性

図 3 と図 4 の比較によって得られる時空間的局所性について述べる．rescue というテーマに着目する．チリにおける rescue というテーマの ValidTime は，(2010-08-05...2010-08-11)，(2010-10-07...2010-10-20) である．中国における rescue というテーマの ValidTime は，(2010-08-19...2010-09-01)，(2010-10-07...2010-11-03) である．さらにここからテーマ rescue の ValidTime と ValidSpace を求めると，((チ

図 5　情報推薦を行うアプリケーションの例

リ), (2010-08-05...2010-08-11)), ((中国), (2010-08-19...2010-09-01)), ((チリ, 中国), (2010-10-07...2010-10-20)), ((中国), (2010-10-21...2010-11-03)) となる．この結果から，rescue というテーマの時空間的局所性が得られた．

5. 考　　察

4. 節で用いたチリと中国の各テーマの出現確率から，テーマが出現する時間や地域には違いがあることを示した．この結果から，ニュース情報は伝播している内容に時空間的局所性が存在する事を意味する．また，このようにして，チリや中国だけでなく，日本やインドなどの各地域や各テーマの比較を行うことで，時空間的局所性が存在するテーマの自動検出が可能になる．さらに，時空間情報を用いることで，局所性が存在する時間や空間を特定できることが分かる．一方で，それぞれのテーマが冗長であったり，テーマだけを参照しても，何について書かれているか不明確であるテーマが存在しており，ユーザに推薦するためには適していない結果となっている．この原因として，はテーマを各文書のタイトルと本文の一文目に出現する単語として用いたことが考えられる．テーマを出現する単語として扱ったため，テーマの時空間的局所性を比較する際に，類似するテーマであっても異なるテーマとして判別されてしまっている．これを解決するために，今回用いたテーマ集合から，被災状況に関するテーマや救済方法に関するテーマなどのカテゴリ分けを行うことが考えられる．これにより精度の高い時空間的局所性が得られると考えられ，また，そのカテゴリ内で，時空間的局所性を比較することによって，同カテゴリであるが，地域や期間によって語られていないテーマなどの局所性の検出が可能になると考える．

6. アプリケーション例

本研究によって実現されるアプリケーションの例を図 5 に示す．図 5 の左側に表示されているのがユーザが現在参照しているニュース情報であり，右側に表示されているのが，システムがユーザに推薦しているニュース情報へのリンクとする．

このシステムでは，ユーザが現在閲覧しているページに関するテーマ集合の中で，閲覧しているニュース情報と同一のテーマであるが，情報の内容が異なるような情報を推薦する．ユーザが他のテーマに興味を持った場合，テーマ名をクリックすることで，そのテーマに関する話題のニュース情報へのリンクの一覧を提示する．ユーザがそのテーマについて深く知りたい場合，現在参照しているページと類似しているページを参照することで情報を得ることが可能となり，そのテーマの全体像を知りたい場合，現在参照しているページとは類似していないページを参照することで情報を得ることを可能とさせる．また，2.2 節で述べた分類の記事数やリンクを示すことで，ユーザはそのニュース情報の伝播に基づいたニュース情報を得ることを可能とする．これによりユーザは，災害などのニュース情報において，被災地域と離れた時空間で語られている同一の話題や，被災地域と近隣した時空間で語られている異なる話題などの情報を網羅的に得ることが出来る．

6.1 用いる手法

このアプリケーションを実現するために，以下の二つの課題を解決する．

- ニュース情報間が異なる情報であるかの判定
- 2.2 節で用いた情報の分類

ニュース情報間で，それらが異なる情報であるかどうか判定を行うために，そのニュース情報が時空間的にどのように情報伝播したのかという情報を用いる．時空間的にどのように情報伝播したのかという情報のクラスタリングを行うことによって，同じテーマであって同じクラスタに属するニュース情報を類似したニュース情報，同じテーマであるが，異なるクラスタに属するニュース情報を類似しないニュース情報として扱う．このようにすることで，ニュース情報の ValidTime と ValidSpace を求め，それらが異なるものを時空間的局所性が存在するとニュース情報として推薦を行う．また，ここで求めたニュース情報間の関係に加え，二次元地図上の直線距離を空間距離，記事の生成された時間の差異を時間距離として算出することによって，2.2 節で用いた情報の分類を行う．

7. 今後の課題

本稿では，2. 節で述べた情報伝播の網羅性について焦点を当てたが，次の課題の考慮が挙げられる．今回の予備実験では，情報伝播の時空間的局所性として，(2) 式で示される，テーマの時間変化による地域間の出現確率を用いたが，(3) 式で示されるテーマの地域変化による時刻間の出現確率を用いることが考えられる．(3) 式を考慮することで，特定の時間において，あるテーマの出現確率の地域的な違いを考慮することが可能になり，それらの変化を調べることで，情報がどのように伝播していったのかを考慮できると考えられる．このように空間的な情報伝播を求めることで，情報伝播の局所性で用いた ValidSpace の正確性を向上させることが出来る．また，3.2 の手法に対しては，今回実験を行えなかったため，3.1 の手法と 3.2 の手法で得られる時空間的局所性の比較を行うことで，どちらの方が有用な手法であるか検証する必要がある．また，2. 節で述べた課題のうち，今回はプリセットとして与えた課題について考慮する必要がある．一般に，各地域に住むユーザは，主に母国語で記載されたニュースを閲覧する．そのため，今回の予備実験で

対象としたチリや日本，中国といった英語以外を母国語とする地域で，母国語で記載されたニュース情報と英語で記載されたニュース情報の特徴に違いが発生している可能性が挙げられる．そのため，言語の違いを超えるために，機械翻訳などを用いた自然言語処理を行うことによって，文書のテーマの抽出を行うアプローチが考えられる．また今回は情報源として，各地域につき 1 つのニュースメディアを予め指定した．しかし，実際にはその地域に情報が伝播されているのにも関わらず，対象としたニュースメディアにのみ伝播していないような情報があった場合に，その地域として情報が伝播していないとみなしてしまう問題がある．また，ニュース情報はニュース記事だけでなく，ブログや Twitter などの情報源も考えられ，これらのニュース情報に対しても時空間的局所性の分析対象とすることが考えられる．このように様々な情報源を扱う際に，ニュース情報が指す時間や空間をどのように特定するかという課題を考慮する必要がある．また，今回はチリの鉱山事故という世界的に関心を集めた災害という話題での分析を行った．しかしながら，世界経済に関する話題や政治に関する話題など，災害とは異なる話題での分析は行っていない．そのため，これらの話題の違いによっても異なる時空間的局所性が得られると考えられる．

8. 関 連 研 究

時間の変化による話題遷移の抽出手法として, [1] や [2] による手法が挙げられる. [1] では，ブログ記事を用いて，ブログ全体の傾向だけでなく，ブログで書かれたある分野をカテゴリとして扱い，カテゴリ内での話題変遷の特徴をトピックを用いて表現し，そのカテゴリ内でもっとも注目を浴びているようなトピックの抽出を目的としている．カテゴリ内で時間ごとに変化するトピックの変化に着目することで，トピックの変化が激しいなどのカテゴリの特徴を抽出し，関連性の高い複数のトピックに関する話題の変遷パターンを発見している．また, [2] ではニュースなどの話題や番組内容が簡潔にまとめられた EPG を用いて，ある話題をキーワードした場合の時間の変化によるサブキーワードの抽出することにより，話題がどのように推移していったのかを抽出し，ユーザの話題把握支援を行う研究である．これらの研究では共に時間の違いによる話題変化の特徴抽出を行っているが，話題の時間的局所性を抽出しているのみで，地域ごとによる空間的局所性について考慮されていない点で本研究とは異なる．

地域の変化による話題遷移の抽出手法として, [3] や [4] による手法が挙げられる. [3] では，サーチエンジンのクエリログを用いて，検索クエリには地域的な特徴が存在することを示している．検索クエリとその IP アドレスから，ある地域においてよく検索されているクエリを抽出し，それらを地図上にマッピングすることで，台風といった検索クエリの中心となっている場所を時間で追うことで，台風の軌跡を辿ったり，野球というクエリに対して，それぞれの地域に所在するチームに関するトピックに関心があるといった地域的特徴を抽出している．この研究では，台風というクエリ例では，時間ごとによって話題が中心となっている地域の抽出を行っているが，本研究では，話題が中心となっている地域以外の情報の特徴も抽出する点で異なる．また，野球というクエリ例では，話題の中心となっている地域を抽出することで，地域の違いによる話題変化の特徴抽出を行っているが，時間変化による時間的局所性に考慮されていない点で本研究とは異なる．また, [4] では，チェンジマイニングの視点から，伝播している情報をクラスタリングした際に，それらのクラスタは時間の変化によって，move したり，grow したりする特徴があることを挙げている．この研究では，このようなクラスタの変化の特徴を挙げているのみであり，本研究では，これらの特徴をモデル化することで，どのように情報が伝播したのかを抽出し，情報伝播の時空間的局所性の検出に利用することを目的とする．

話題の時空間的な特徴を抽出する手法として, [5] や [6], [7] による手法が挙げられる. [5] では，ブログ記事を用いて，ある話題のトピックとなるテーマの時空間パターン検出を行っている．この研究では，本研究と同様に，ある話題のトピックの出現確率を用いることで，時空間パターンの検出を行い，ある地域の時間変化による，各テーマの出現確率の変化やある時間の地域変化による，各テーマの出現確率の変化を図示している．そのため，あるテーマの時間的変化による違いと空間的変化による違いをそれぞれ求めているが，本研究では，時間的変化と空間的変化を同時に求めることで，テーマの時空間的局所性の検出を目標とする点で，本研究とは異なる. [6] では，Twitter を用いて、地震や台風などのリアルタイムイベントの発生時間や地域を即座に感知することを目的としている. [3] の研究でも触れられていた通り，時間ごとに変化する話題の中心となる地域を抽出することで，地震の中心や台風の軌跡を検出しているが，本研究では，ある話題の中心となっている時間や地域だけでなく，その話題のトピックに着目することで，地震の中心や台風の軌跡に関する情報を抽出するだけでなく，地震によって引き起こされた二次災害などのトピックに関する情報の時空間的な特徴を抽出する点で本研究とは異なる. [7] では，時空間データを二次元の地図に時間情報を加えた三次元上にマッピングを行い，その中から興味が類似したユーザをグルーピングするツールを紹介している．三次元上に視覚化したグループは幾つかの形に分類できることを示している．これにより，伝播している情報のグループの形によって，伝播している情報に時空間的な特徴があることを示しているが，それら情報伝播の時空間的局所性の分析を行い，情報推薦に用いる点で，本研究とは異なる．

9. ま と め

本稿では，特定の話題に対するニュース情報の主題となるテーマに時空間的な局所性が存在することを検証するために，ニュース情報から，そのニュース情報が生成された時間と空間，そのテーマを抽出することで，各地域や各時刻における各テーマの出現確率を求め，得られた確率分布を比較する手法を提案した．また，この手法の有用性を評価するために，世界的な関心を集めたチリの鉱山事故に関するニュース記事を用いて予備実験を行うことで，チリの鉱山事故という話題に対するテーマの時空間的局所性が存在することを示した．また，これらの結

果で時空間局所性をユーザに推薦するためのアプリケーション例を提示することで，本研究の有用性を示した．

謝辞　本研究の一部は，京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」，および，文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発時代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」（研究代表者：田中克己，A01-00-02，課題番号：18049041）によるものです．ここに記して謝意を表します．

文　　献

[1] 戸田 智子, 福田 直樹, 石川 博, “ Blog 記事のクラスタリングに基づいたカテゴリ別話題変遷パタンの抽出 ”, DEWS2007, A8-Blog, 2007

[2] 菊池 匡晃, 岡本 昌之, 山崎 智弘. 階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出. データ工学ワークショップ DEWS, (2008).

[3] Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. 2008. Spatial variation in search engine queries. In Proceeding of the 17th international conference on World Wide Web (WWW '08). ACM, New York, NY, USA, 357-366.

[4] Mirko Böttcher, Frank Höppner, and Myra Spiliopoulou. 2008. On exploiting the power of time in data mining. SIGKDD Explor. Newsl. 10, 2 (December 2008), 3-11.

[5] Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In Proceedings of the 15th international conference on World Wide Web (WWW '06). ACM, New York, NY, USA, 533-542.

[6] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 851-860.

[7] Kyoung-Sook Kim, Koji Zettsu, Yutaka Kidawara, Yasushi Kiyoki, "StickViz: A New Visualization Tool for Phenomenon-Based k-Neighbors Searches in Geosocial Networking Services," Conference, International Asia-Pacific Web, pp. 22-28, 2010 12th International Asia-Pacific Web Conference, 2010