

TRENDREADER(TR)を用いた災害・危機に関する言語資料体の 解析にもとづく災害・危機事象の将来展開予測手法の開発 - 新潟県を襲った近年の2つの地震災害を例にして -

The Development of a Disaster Social Phenomena Trend Prediction Method
Based on Analysis of Corpora about Disasters and Crises Using TRENDREADER™
- A Case Study of Web News Corpora for the 2004 Niigata-ken Chuetsu Earthquake
and the 2007 Niigata-ken Chuetsu-oki Earthquake -

佐藤 翔輔¹, 林 春男²

Shosuke SATO¹ and Haruo HAYASHI²

¹ 京都大学大学院 情報学研究科

Graduate School of Informatics, Kyoto University

² 京都大学 防災研究所

Disaster Prevention Research Institute, Kyoto University

Emergency Operation Center should have intelligence for prediction not only about probability and scale of natural hazard, but also tendency of social phenomena of disaster for proactive disaster response. However, unlike natural hazard estimation, there is no method or system that could predict social trend. In this paper, we propose a prediction method of social phenomena tendency based on analysis of disaster corpus using TRENDREADER(TR). The TR system automatically extracts meaningful keyword sets from digital text datasets related to disasters, and visualizes a “big picture” of whole situation. Our method consists of the following two steps; 1) to identify keywords among documents about possibly significant disaster social phenomena and 2) to apply trend of the keywords in stock chart model and to provide daily summary as well as weekly summary.

Keywords: disaster social phenomenon, proactive disaster response, disaster prediction, text mining, keyword extraction, stock chart

1. はじめに

危機対応を効果的に行うためには、ハザード・被害・対応などの状況認識の統一（Common Operational Picture: COP）を的確に図る必要があるほか¹⁾、プロアクティブな危機対応（先見的な危機対応）を行う上では、状況の展開を予測できることが望ましい。

「自然現象としての災害」を取り扱う領域においては、余震の発生確率や台風の進路や規模の予測など、自然ハザード現象の予測の技術が発展し、実際の災害対応の場面においても活用されている。一方、「社会現象としての災害」を取り扱う領域では、災害・危機が発生しているときに中長期的に被災者にふりかかる問題や対応課題の需要の大きさの変化などを予測・情報提供するような仕組みは存在しない。危機管理の実務者は、平時において擬似的な災害・危機の環境下で、危機対応行動を行う図上訓練を実施することで、災害のイメージを構築し、実際に災害・危機が起きた場面での事態予測に結びつけようと努力している。また、近年の災害では、被災県の

災害対策本部会議に対応経験のある実務者や大学等の研究機関に所属する危機管理の専門家が同席し、これまでの経験にもとづいた事態の展開の見通しと、とるべき対応方針を助言するような試みも見られる。前者は、予測できる状況のバリエーションが、訓練の事例となった擬似的な災害の特性や状況付与内容の質に大きく制約される。後者は、今世紀前半に発生が確実視されている東海・東南海地震津波のように、広域にわたって甚大な被害をもたらすような事態において、経験のある実務者や専門家の人数が限られているような状況下では、広範にわたる被災地への専門家の投入は期待できない。

「社会現象としての災害」の展開の予測は、佐藤ら²⁾がこれまでに開発してきた災害・危機に関する言語データからキーワードを自動抽出するシステム（TRENDREADER:TR, トレンドリーダー）を応用することで実現できる可能性がある。先行研究で開発した手法は、災害の社会現象の側面を記録するデータの形態が、主に言語データであるために、収集・整理・分析・閲覧などが困難であるという問題を解決するため、災害・危

機において事象が展開する状況を概観することをねらいとして、災害・危機に関する言語資料体（コーパス）からキーワードを自動抽出し、時系列的に可視化することができる。2007年地域安全学会研究発表会（秋季）では、佐藤らの提案システムが報告された上で、フロアから「この手法を使って災害事象の展開を予測することができないか」という指摘がなされた。また、これまでも危機管理に関連する研究報告会等において、上記と同じような質疑が投げかけられている。この手法は、時系列的に増加するコーパス中の言葉について、任意の時間断面における特異性を数値的に評価し、時系列のデータセットを生成するものである。この時系列データの傾向変動を的確に捉えることができれば、任意の災害・危機事象の展開を予測できる可能性がある。

災害・危機の将来展開を予測しようとするさいには、災害下の現実を適切に把握したうえで、事象の兆候を分析する必要がある。人が現実を把握することは、現実の一部を操作化し、可視化された表象をつくることから始まる。この表象を複数の人々が共有・認識することで形成される社会的現実しか、人は現実を知りようがない。現実が多面的で複合的であるために、複数の人が現実を言語、映像、音声、数式など様々な表象に置き換えようとする。現実を把握することは断片的な表象を体系的に統合することが前提であるが、以上の中でも言語は、比較的統合が容易な表象であるほか、社会的な現象を記述することに優れている。A. ハートマン⁴⁾は「言葉が世界を創る（Words create worlds）」と述べていることから、言語データを現実の近似値として位置づけることが有効であることが分かる。この研究では、以上にもとづき、災害下の社会的現実を記述した言語データを解析することにより、将来展開の予測を試みることを考える。

この研究では、プロアクティブな危機対応を支援するために、災害や危機が発生しているさなかに、今後の災害・危機事象がどのように展開するのかを予測する手法を開発することを目的とする。本稿では、佐藤ら²⁾が開発した災害・危機に関する言語データ群から、時間断面ごとに言葉の特異性を評価し、トレンドとして可視化する手法を応用することで、上記の目的を検討する。

先行研究を応用して上記の目的を実現するアプローチには、次の2つのパターンが考えられる：1) 対応に当たっている当該の災害・危機と、それに類似する地域性や同一のハザード種に起因したイベントについて、同一のキーワードの時系列的なふるまいを比較する方法、2) 時系列データの変動傾向にもとづき、動向予測を行う既存の手法を援用する方法。前者は、災害と災害のもつ共通性・普遍性に着目し、同一の事象は時間的・量的にほぼ同じような展開を見せるという仮定にたったアプローチである。後者は、すでに研究が進んでいる株取引分野に見られる株価動向予測の手法や考え方を活用しようとするものである。

4章では1)のアプローチを、5章では2)の方法で災害・危機事象の将来展開予測手法の開発を試みる。具体的には、4章で1)のアプローチによる将来展開予測の限界を示し、5章で2)の有用性を述べた上で、ここでの検討にもとづく災害・危機事象の将来展開手法の手順設計を6章で行う。

2. TRENDREADERの概要

ここでは、先行研究で開発した災害・危機に関する言

語資料体（コーパス）から、災害過程が時々刻々と変化する様子をキーワードを用いて可視化し、大量の文書がもつ情報を概観するTR²⁾の概要を述べる。

この手法は、TFIDF⁵⁾とIncremental IDF⁶⁾の考え方を組み合わせた時間増加型TFIDF²⁾という語の重み付け指標を計算し、キーワード自動抽出を試みるものである。災害や危機に関する言語データは、発生からの時間の経過に伴って、データ量が徐々に増加していく。通常のTFIDFには時間パラメータが存在しないために、このようなコーパスを対象にした重み付け計算は行えない。一方、Incremental IDFは時間経過ごとに異なる総文書数を変数としている。以上2つの指標を組み合わせ、時間の概念を考慮した語の重み付け指標として以下の式で表される時間増加型TFIDFを提案した：

$$\text{時間増加型TFIDF}(t_i, d_j) = TF(t_i, d_j) \cdot \log \frac{N_j}{DF(t_i, d_j)} \quad [1]$$

t_i ：時間増加型TFIDF(t_i, d_j)を算出する対象となる単語。 i はその識別子を表す添え字。

d_j ：時間増加型TFIDF(t_i, d_j)を算出する単語 t_i が含まれている単位文書。文書の単位は文章、段落、文など任意に定める。この研究では、記事を文書単位としている。 j は識別子を表す添え字。

N_j ：文書 d_j が出現した時点での文書の総数。

$TF(t_i, d_j)$ ：Term Frequencyの略。単語 t_i が文書 d_j に出現した回数。

$DF(t_i)$ ：Document Frequencyの略。キーワード t_i が出現する文書数（文書頻度）。

テキストデータベースからテキスト群を形態素解析⁷⁾によって形態素への分かち書きをおこない、品詞情報にもとづいて、非自立と接尾以外の名詞・動詞・副詞・形容詞を抽出する。抽出した形態素について、単位文書（記事）ごとに、単語の出現頻度TFと、時間増加型TFIDFを算出する。

TRの手法では、任意時間幅に増加したテキストデータ群の中からより特異的な語を抽出することをねらいとしている。これを実現するために、任意の時間幅だけさかのぼったコーパスの特性と現時点でのコーパスの特性を比較し、その差異を現時点における特異性の度合いとし

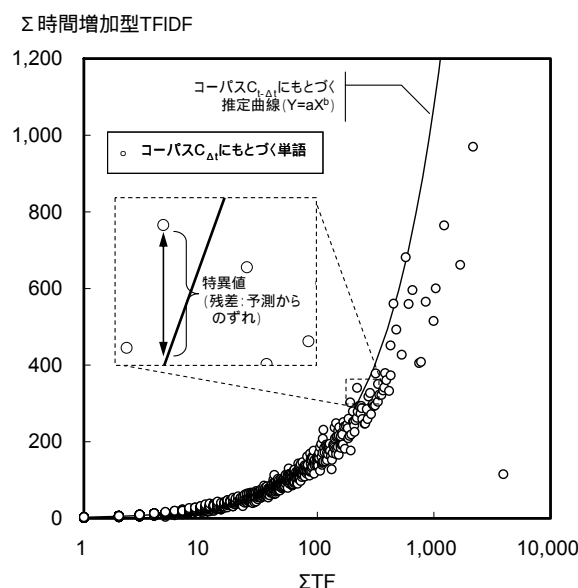


図1 TRによるキーワード自動抽出
(ΣTFとΣ時間増加型TFIDFの関係)

て捉えている。

ある時点 $t-\Delta t$ から t の間 Δt におけるコーパス $C_{\Delta t}$ (コーパスの増分) からキーワードを抽出するために、時点 $t-\Delta t$ までのコーパスにもとづく TF の累積値 (ΣTF) と時間増加型 $TFIDF$ の累積値 (Σ 時間増加型 $TFIDF$) の関係式を求める。佐藤らは、 ΣTF と Σ 時間増加型 $TFIDF$ の間には、強い曲線関係があることを明らかにした (図 1)。これを受け、 ΣTF を独立変数 X 、 Σ 時間増加型 $TFIDF$ を従属変数 Y にした $Y=aX^b$ (a, b は定数) の曲線のあてはめをおこなった推定式を用いて、時点 t の ΣTF を入力値とした時点 t の Σ 時間増加型 $TFIDF$ の推定値を求める。時点 t で実際に計測された Σ 時間増加型 $TFIDF$ の実測値から、ある時点 t から Δt さかのぼった時点のコーパス $C_{t-\Delta t}$ の特性から予測される Σ 時間増加型 $TFIDF$ の推定値を減じたもの (残差値) を、過去と現在の言語分布特性のずれとして捉え、単語の特異性として数量的に評価する (図 1)。この残差値を特異値と呼んでいる。TR は、横軸を日時、縦軸を特異値としたグラフに言葉 (キーワード) をプロットし、キーワードの時系列的推移を可視化する。

3. 予測手法の適用対象の選定と収集

この研究では、社会現象としての災害・危機事象の将来展開予測手法の検討や構築、適用を試みるためのコーパスとして、2004 年新潟県中越地震 (以下、中越地震) と 2007 年新潟県中越沖地震 (以下、中越沖地震) について報道されたウェブニュース記事を用いることにする。災害・危機についてのウェブ報道は発災から継続的に配信されることと、デジタル媒体であることから、応急対応や復旧・復興までを包括する内容の言語データを比較的容易に蓄積・加工できるという特徴がある。

上記の 2 つの災害を対象にしたのは、ハザードが同じであることと、同じ県域内で起こった災害であるためである。1 章で述べたように、災害事象の将来展開を予測するアプローチには 2 種類が考えられ、そのうち 1 種類は、ハザード種や地域性の類似する当該の災害と過去の災害のキーワードのふるまいを比較しようとするものである。このようなアプローチを用いた手法を開発し、その有用性を検証するためには、同じ新潟県で発生した 2 つの地震災害の例を取り上げることが最も妥当であると考えた。

図 2 に両地震について代表的なポータルサイト⁸⁾ 上の「新潟県中越地震」「平成 19 年新潟県中越沖地震」というトピックス中にアップデートされたウェブニュースを自動収集した記事について、時間経過にともなう累積記事数を示した。両地震について、発災から 4,000 時間 (約半年) 収集したニュース記事は、中越地震で 2,588 件、中越沖地震で 3,042 件であった。ニュース記事数について、中越沖地震が中越地震を上回るものの、記事数の積み上がり方はほぼ同じような傾向を示していることから (図 2)、2 つの地震災害はハザードや地域性のみならず、社会の関心の度合いもほぼ同じ程度であったことが分かる。

以降では、上記の 2 つの地震災害のウェブニュース記事を用いて、災害・危機事象の将来展開予測手法の開発・検討を進めていく。

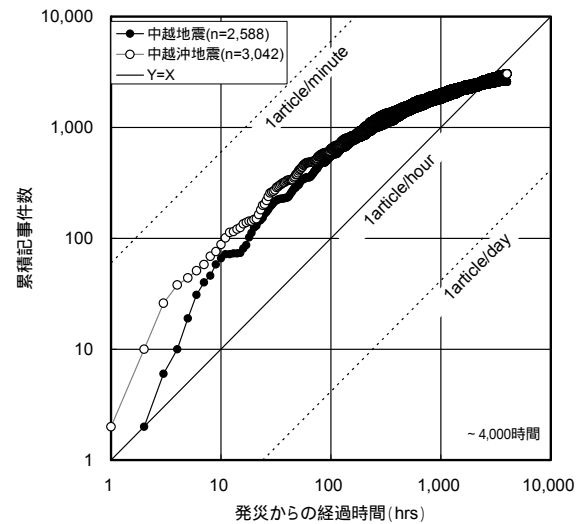


図 2 中越地震と中越沖地震について報道された Yahoo!ニュース記事数の時系列的推移 (発災から 4,000 時間まで)

4. 類似する 2 つの災害に関するコーパスから抽出した同一のキーワードのふるまいにもとづく将来展開予測手法の検討

複数の災害・危機に着目するとき、ハザードが同一であったり、その規模や地域性が類似するような場合には、発災後に起きる重要な事象の内容やタイミング、需要の大きさなども似たような傾向を示すと考えられる。TR の手法では、キーワード抽出結果から単一のキーワードを取り出すことで、任意のキーワードの特異値の変動 (ふるまい) をグラフ化することができる。異なる 2 つの災害の同一キーワードを同時に図示すれば、キーワードのふるまい、ひいてはキーワードに関連する災害事象のトレンドを比較することが可能になる。

この章では、前記の仮説のもと、中越地震と中越沖地震のウェブニュースから、1) TR の手法にもとづいて重要なキーワードを同定し、2) 2 つの地震について 1) で求めたキーワードの内容と、3) 同一キーワードの特異値のふるまいを比較する。これらの分析を経ることにより、両地震において重要な災害事象の同一性、普遍性の有無を確認し、同一キーワードのふるまいを比較することによる将来展開の予測の可能性を確認する。

(1) 重要キーワードの同定

任意のキーワードを闇雲に選び、将来動向を予測しようとしても、選んだキーワードに関連する災害事象が当該の災害において重要でなければ、予測の効果は低い。ここでは、中越地震と中越沖地震で重要だった事象に関連するキーワードを明らかにし、同時に重要なキーワードを同定する手法を開発する。

TR の手法では、特異値という指標で任意の時間断面における特異性を評価する。この特異値を積み上げることで、積み上げを終えた時点までの各キーワードの重要性を評価できると考えた。発災から継続的にキーワードとして抽出された言葉に関連する事象は、将来展開の予測すべき対象であると考えられる。

具体的には、特異値 (Discriminating Value) を発災から任意の時間断面まで足しあわせた数値を計算する、こ

の指標を、累積特異値 (ΣD) と呼ぶことにする。特異値が正を示した言葉は各時間断面を特徴づける言葉で (特異語)、負の値を示した言葉はコーパスに遍在する言葉である (共通語)。単純に特異値を足しあわせれば、任意の言葉が特異語から共通語になった場合、負の値が足し合わされていく。このような言葉は、ある一定の期間に重要な事象に関連するキーワードであったにもかかわらず、積み上げられた特異値が減じられることになってしまい、重要なキーワードとして求めることができない可能性がある。ここでは、正の特異値だけを足し併せたものを累積特異値とする。

青野ら⁹⁾と田中ら¹⁰⁾によって、1995年の阪神・淡路大震災の被災者の発災直後からの行動についてミクロな視点からエスノグラフィーが丹念に採取され、災害過程がモデル化されている。これによれば、災害過程において時間は、10時間、100時間、1,000時間と10のべき乗の時間によって状況が質的に変化するとされている。2つの地震災害のウェブニュースコーパスについて、10時間までの1時間刻み、10~100時間を1時間刻み100~1,000時間を24時間刻み、1,000~4,000時間を24時間刻みで Δt を設定し、TRの手法を適用し、発災から10時間後、100時間後、1,000時間後、4,000時間後の累積特異値を求め、重要キーワードを同定することにした。

累積特異値を求めた結果と、それにもとづく重要キーワードは(2)節で記述する。

(2) 重要キーワードの比較

各時間断面 (10時間、100時間、1,000時間、4,000時間) ごとに、2つの地震で共通するキーワードについて、累積特異値を比較する。これは、キーワードの内容や意味の視点から、2つの地震災害で共通して重要となったキーワードや、どちらか一方の地震災害を特徴づけたキーワードを把握することをねらいとしている。

上記のような判断を補助するために、累積特異値にカットオフ値を設けることを考えた。設定したカットオフ値が統計的に有意な値であれば、重要なキーワードが否かを客観的に判断することができる。ここでは、重要なキーワードを判別するカットオフ値として、累積特異値分布の上位5%の累積特異値を採用することにした。5%という値は、統計学において有意水準として用いられている値であり、異常に大きな数値と、そうでない数値を判別する基準として用いることができる。

図3に2つの地震災害のウェブニュースから抽出したキーワードのうち、両地震で出現したキーワードの累積特異値の関係を示した。図3には、累積特異値分布の上位5%のカットオフ値を破線で示している。両地震で累積特異値がカットオフ値を超えたものは、2つの地震災害で共通する重要なキーワードであり、どちらか一方の地震災害でカットオフ値を超えたキーワードは、当該の地震災害で特徴的なキーワードであることを示している。

10時間までは「余震」が両地震で重要なキーワードだった。中越地震では「断層」「電話」「死亡」というキーワード、中越沖地震では「ガス」「家屋」「現地」というキーワードが特徴的であった。ここでは、ハザードや直接的な被害に関するキーワードが多く見られる。

100時間までは、共通する重要なキーワードには「ガル」という地震の加速度を表す単位、「ボランティア」「電話」などがある。中越地震で特徴的なのは、「IC (インターチェンジ)」「レール」といった交通系のライフラインの被害に関連するキーワード、中越沖地震で

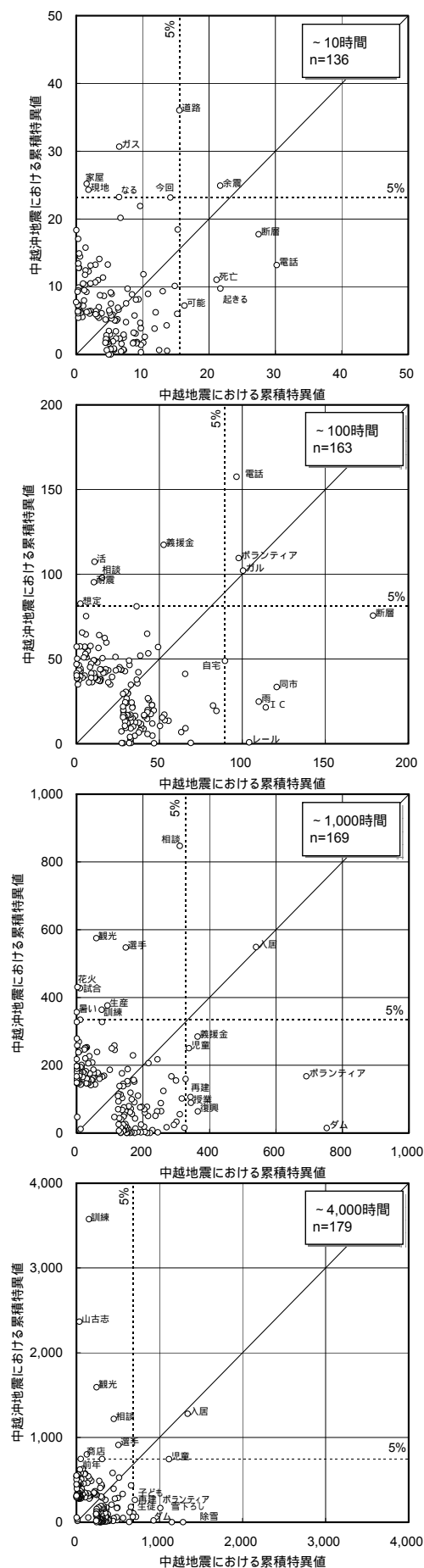


図3 中越地震と中越沖地震から抽出されたキーワードの累積特異値の関係

は「義援金」「相談」などのほか、柏崎・刈羽原子力発電所の被災に関する「活（活断層）」「耐震」「想定（被害想定）」のようなキーワードが特徴的だった、

1,000 時間までは、 $Y=X$ の線の上に「入居」が示された。中越地震で特徴的だったのは「ボランティア」「ダム（天然ダム）」、さらに「再建」「復興」も見られる。中越沖地震は、原発事故による風評被害が発生したことを受けた「観光（地元の観光産業）」、生産工場の活動停止を受けた「生産」というキーワードが特徴的だった。

4,000 時間まででも「入居」は両地震でほぼ同じ累積特異値を示した。中越地震では「ボランティア」「除雪」というキーワード、中越沖では「観光」「相談」というキーワードが特徴的である。後者に「山古志」とあるが、中越沖地震のニュース記事の中、同じ地域の過去の地震災害の被災地の復興はどのようにになっているかについて報じた記事が含まれていたことが大きな原因になっている。

10 時間、100 時間、1,000 時間、4,000 時間のそれぞれで累積特異値が著しく高かったキーワードは、時間の経過にともなって、命を守る活動、社会のフローを回復させる活動、社会のストックの活動¹¹⁾に関連するものが見られたほか、中越地震では芋川の河道閉塞（「ダム（天然ダム）」、雪害対応（「除雪」「雪下ろし」）、中越沖地震では原発事故（「活（活断層）」「耐震」）、風評被害（「観光」）、地域産業の停止（「生産」）などの事象が表れた。このことは、累積特異値を計算し、上位 5% の値をカットオフ値として用いることで、コーパス中で着目すべき、予測する価値がある重要キーワードを同定できていることを示している。

以上で示された重要キーワードを表 1 に整理した。表 1 には、時間断面ごとに両地震で累積特異値がカットオフ値を超えたキーワード、どちらか一方の地震でカットオフ値を超えたキーワードに分けて整理している（表 1 の上部）。さらに、これらのキーワードを単独の時間断面で重要キーワードになったもの（単独フェイズ）、複数の時間断面で重要キーワードになったもの（複数フェイズ）、2 つの地震で重要キーワードとなった時間がずれたもの（フェイズにずれ）に再区分したものも示している（表 1 の下部）。

両方の地震災害において、同じフェイズで重要なキーワードになった例は少ない（表 1 の下部）。これらの例は、複数のフェイズにまたがって出現した「入居」と、発災から 100 時間以内に重要キーワードになった「余震」「道路」「ガル」のみである。この結果はハザードが同一であり、時代変化と地域性の異なりが比較的小さい災害であっても、重要になる災害事象の内容とそのタイミングが異なることを示している。ハザードや地域性が類似する過去の災害のキーワードを参照しても、当該の災害の重要な事象を予測できない可能性が示唆された。

(3) 重要キーワードのふるまいの比較

災害事象の将来展開を予測する上では、重要となる事象を察知するだけでなく、重要になるタイミングや量的な情報が必要になると考えた。(2)節で同定された重要キーワードについて、TR の手法で求められる任意キーワードの特異値の時系列データセットを準備し、2 つの地震における同一のキーワードのふるまいを比較し、災害事象の将来展開予測の可能性について検討する。

特異値の時系列データをグラフ化する上で、そのふるまいをより理解しやすいものにするために、移動平均法

表 1 中越地震と中越沖地震における重要キーワード

時間	累積特異値がカットオフ値を超えたキーワード（重要キーワード）		
	両地震	中越地震のみ	中越沖地震のみ
～ 10	余震、道路	電話、断層、起きる、死亡、可能	ガス、家屋、現地、なる
～ 100	電話、ボランティア、ガル	断層、同市、IC、雨、レール	義援金、活、相談、耐震、想定
～ 1,000	入居	ダム、ボランティア、復興、義援金、授業、再建、児童	相談、観光、選手、試合、花火、生産、暑い、訓練
～ 4,000	入居	除雪、雪下ろし、児童、ボランティア、ダム、生徒、子ども	訓練、山古志、観光、相談、選手、商店、前年
フェイズにずれ	電話、義援金、ボランティア		
複数フェイズ	入居	断層、ダム、児童	相談、観光、選手、訓練
単独フェイズ	余震、道路、ガル	起きる、死亡、可能、IC、雨、レール、復興、授業、再建、除雪、雪下ろし、子ども	ガス、家屋、現地、なる、活、耐震、想定、花火、試合、生産、暑い、山古志、商店、前年

を用いてデータの平滑曲線化を行った。移動平均値を算出する項数は、一般的に用いられている、3 項、5 項、7 項を採用した。

紙面の制約上、表 1 の整理結果にもとづき、両方の地震災害が同じタイミングで重要キーワードになった「入居」（社会のストックを再建させる活動の一例）と、100 時間で両地震、1,000 時間で中越地震のみで重要キーワードになった「ボランティア」（社会のフローを安定させる活動の一例）、中越沖地震で重要キーワードになった「観光」について、特異値のふるまいの比較結果を図示し、観察を試みる。また、これ以外に命を守る活動に関連するキーワードの例として「死亡」も図示・観察した。

「死亡」の中越地震と中越沖地震における特異値のふるまいは、どちらも 10～100 時間の間に特異値のピークを迎え、100 時間以内に収束していく傾向を示した（図 4）。この結果は、災害救助における”Golden 72 hours”と適合している。

図 4 に「入居」について、両地震災害における特異値のふるまい（特異値と移動平均値）を示した。「入居」は、主に「応急仮設住宅への入居」に関連する記事から抽出されたキーワードである。「入居」は、発災から 100 時間頃あたりから特異値が両地震で同じような値で増加し、およそ 1,000 時間後（およそ 1 ヶ月半後）にピークを迎え、その後も同じような値で低下していった。「入居」は図 3 の 1,000 時間、4,000 時間において累積特異値が中越地震と中越沖地震でほぼ同じ値を示したことから、両地震で共通する重要な事象に関連するキーワードであり、重要になるタイミングも似たような傾向を示すことが予想された。ある時間断面で累積特異値が近い値を示した言葉については、特異値のふるまいが似たような傾向を示す例があり、キーワードに関連する将来展開の予測が行える可能性が見られた。

一方、図 4 の「ボランティア」と「観光」の特異値は、両地震災害で異なる特異値のふるまいを示した。「ボランティア」は、被災地支援のボランティア活動に関連する記事から抽出されたキーワードであるが、中越地震では、100 時間から 1,000 時間の間に特異値のピークがあったのに対し、中越沖地震では 100 時間付近にピークが見られ、特異値の最大値も中越地震を下回る。中越沖地震のウェブニュース中の記事によれば、発災後の初の土日に被災地を訪れたボランティアは中越地震で 5,666 人（長岡市、小千谷市、十日町市、川口町）、中越地震で 2,296 人（柏崎市、出雲崎町、刈羽村）であり、被災地支

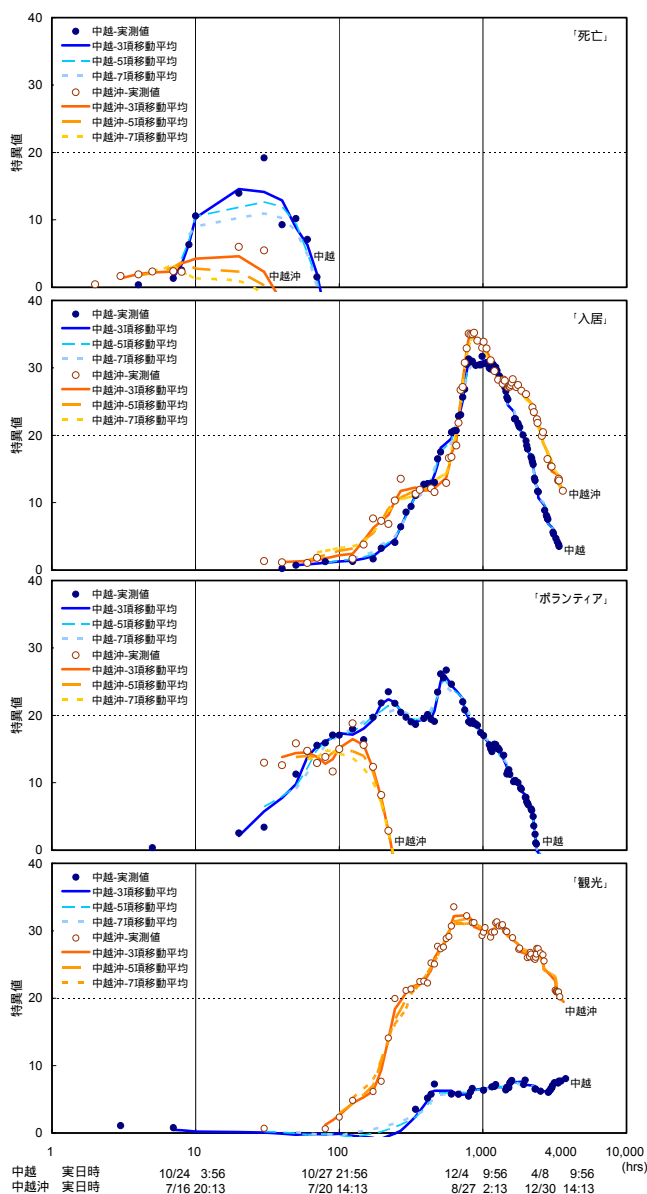


図4 重要キーワードのふるまいの比較
（「死亡」「入居」「ボランティア」「観光」）

援が低調であることが述べられている（7/25 15:06 毎日新聞）。災害対応活動として、一般化・普及が進んでいるボランティア活動という事象であっても、災害間で異なる展開あったことが分かる。また、原発の被災による地域観光産業への風評被害という、中越沖地震において特徴的だった事象は、当該の災害と過去の災害における同一のキーワードは同じ特異値の変動傾向を示さないことは、「観光」のふるまいからも明瞭である。

本章での結論は、着目している災害・危機における事象について、当該の災害と過去に発生したハザードや地域性が類似する災害から抽出された同一のキーワードのふるまいから、将来展開を予測することは困難であるということである。一部、累積特異値がほぼ同じ値を示したキーワードについては、上記を実現できる可能性があるものの、そうではないキーワードの数は多く、本章の流れで予測評価を行った場合には、事実とかけはなれた結果を予想する可能性が高い。

一方で、2つ（複数）の災害や危機について、重要キ

ーワードの累積特異値と、その特異値のふるまいを比較することは、言語データの中から災害の共通性・普遍性や特徴を客観的に分析する手段として有効であることが分かった。ここでの分析の流れは、当該と過去の災害・危機に関するコーパスの特性を定点的、時系列的に比較し、教訓や問題を導く手法として位置づけられる。今後は、以降発生する災害・危機事案について同様の分析を継続し、本稿の分析結果で見られたような「死亡」や「入居」など同じタイミングで繰り返される災害事象を明らかにし、これらを結果を整備していくことにする。

過去の災害の重要キーワードのふるまいを参照しても、着目しているキーワードのふるまい、それに関連する災害事象の展開を予測することは困難であることが分かったことから、次章では、単一の災害・危機に関するコーパスのデータから、その変動傾向を分析し、予測評価につなげる手法について検討する。

5. 株価動向予測手法を援用した災害事象の将来展開予測手法の検討

時系列データの動向を予測するというオペレーションは、株取引において実際に行われている。この分野では損失をいかに減らし、利益をいかに伸ばすという理念のもと、着目する株券の価値が増加するか・減少するかを予測しながら、株取引が行われている。一般にトレーダーは、株価の変動（時系列データ）を日単位や週単位で集約し、集約結果と蓄積された傾向パターンをつきあわせ、増加・減少を予測している。1日間などのオペレーションの周期を定め、それまでの状況进行分析し、予測につなげ、次の対応策を検討するという流れは、责任担当期間（Operational Period: OP）を設け、状況認識の統一（COP）を図り、業務計画（Incident Action Plan）を策定するという危機対応の標準的な流れ¹⁾と酷似している。著者らは、以上のような特徴に着目し、危機対応の実務への取り組みやすさを踏まえ、株取引における株価動向予測方法を災害事象の展開予測に援用することを考えた。

この章では、1) 将来展開を予測すべきキーワードを同定する方法、2) 株価動向分析手法の援用して災害事象の将来展開を予測する手法の開発を試みる。なお、この章では、中越沖地震のウェブニュースコーパスについて、1時間刻み（ $\Delta t=1$ 時間）でTRの手法を適用した結果から求められた特異値のデータセットを用いる。

(1) 将来展開を予測すべきキーワードの同定

4章（1）節で述べたように、災害事象の将来展開を予測しようとする場合に、当該の災害にとって重要な可能性の高い事象を選択しなければ、予測とそれを受けた対応策の効果はおろか、予測の意味そのものが薄れてしまう。この節では、将来展開を予測するにあたりする事象に関連するキーワードを同定する方法について検討する。

前章で、重要なキーワードを同定する方法として、累積特異値（ ΣD ）を用いることと、カットオフ値を設けることに、一応の妥当性があることを確認している。ここでも、将来展開を予測すべき重要なキーワードを評価する指標として累積特異値を採用し、ある基準に従うカットオフ値を与えることを考える。

カットオフ値を与える方法には、1) 上位から語数が

ある割合（1%，5%など）になる累積特異値をカットオフ値にする方法，2）上位から任意の語数（10位，20位など）になる累積特異値をカットオフ値にする方法，3）過去の経験値から定数として与えられた累積特異値をカットオフ値にする方法などがある．

これらの方法をもとに，将来展開を予測すべきキーワードか否かを判別する累積特異値の適切なカットオフ値を検討するため，累積特異値の分布パターンを観察する．累積特異値の分布特性が明らかになれば，カットオフ値を求める適当な方法を定めることができると考えた．図5に中越沖地震から抽出されたキーワードの特異値のデータセットから累積特異値を求め，100時間（約4日），300時間，400時間，500時間（約3週），750時間（1ヶ月間），1,000時間（約6週），1,500時間（2ヶ月），2,000時間，3,000時間（4ヶ月），4,000時間ごとの分布をみた（ヒストグラム）．累積特異値の分布は，低い値で度数が大きく，高い値の度数が少ないロングテール型の分布であったため，累積特異値を常用対数で対数化した数値でグラフ化している．平均値は真数で6.0～17.3（常用対数：0.78～1.24）で，中央値は一貫して真数で1.3（常用対数：0.11）を示し，累積特異値は値の低い側（左側）に歪み，累積特異値が高まるほど度数が減少するという分布を示した．このような分布は，累積特異値の高い付近に，ある基準値を設けることで重要なキーワードの絞り込みが行えることを意味している．

図5には重要キーワードを絞り込む基準の候補になり得るものとして，上位から語数がある割合になる累積特異値をカットオフ値にする方法（前述1））にもとづき，上位1%，5%の累積特異値の値（真数）と，それを境界とすることで絞り込まれる重要なキーワードの数（語数）をかつこ内に示している．どちらも時間の経過に伴って，境界となる累積特異値が増加するものの，言葉の種類（n）も増加するために，選定されるキーワードの数が増えていく．上位1%では，言葉の数がおおよそ200語に収束される傾向が見られるものの，将来展開を予測するに当たり，該当するキーワードが何百とあれば，参照しなければならないキーワードが増えるため，展開予測分析のさいに多くの時間を要し，対応策の意思決定に支障をきたしてしまう．

次に，上位から任意の語数になる累積特異値をカットオフ値を与える方法（前述2））にもとづき，順位（ランク）を用いて，将来展開を予測すべき重要なキーワードを絞り込むことを考える．任意の順位を設定することで，絞り込まれるキーワードの数はこれに制約され，分析すべきキーワードを適度に減らすことができる．

図6には，累積特異値の上位1%や5%，上位10位，20位，30位，40位，50位，100位となる累積特異値（カットオフ値）の時間変化をプロットした．発災からの経過時間とカットオフ値の間には，強い指数曲線（ $Y=a\ln(X)+b$ ）の関係があることが分かった．時間経過にともなって，上位1%や5%および順位にもとづくカットオフ値が系統的な傾向を示したことから，ある言葉について累積特異値を求めたときに，この曲線のどこに位置し，どれだけ重要な言葉かの目安を与える検量曲線となり得る．複数のイベントについて同様の分析を行い，曲線の係数を再計算しつづけることで，累積特異値にもとづくカットオフ値を定数として求めることができる．今後のデータを蓄積することで曲線の安定化を行っていく必要がある．

1）累積特異値の分布から上位1%，5%の値をカットオフ値とする方法，2）順位にもとづくカットオフ値を与える方法，および3）経験にもとづく定数をカットオフ値とする3つの方法を将来展開を予測すべきキーワードを判別するさいのオプションとして整備することにした．上位1%や5%をカットオフ値として採用すれば，上述したように選択されるキーワードの数が増える．一方で，定量尺度の分布における上位1%や5%は，統計学で経験的に用いられたきた実績のある基準である．順位を基準とする方法は，適当な数のキーワードを選定することができるが，上位1%や5%に含まれるキーワードを取りこぼすことになる．また，過去のデータの蓄積にもとづいて設定された定数のカットオフ値は，他の災害・危機の事例の一般像と比較することが可能になる．それぞれのカットオフ値を定める方法のメリットを勘案し，3つのオプションとして将来展開を予測すべきキーワードを同定する手法に取り込むことにした．

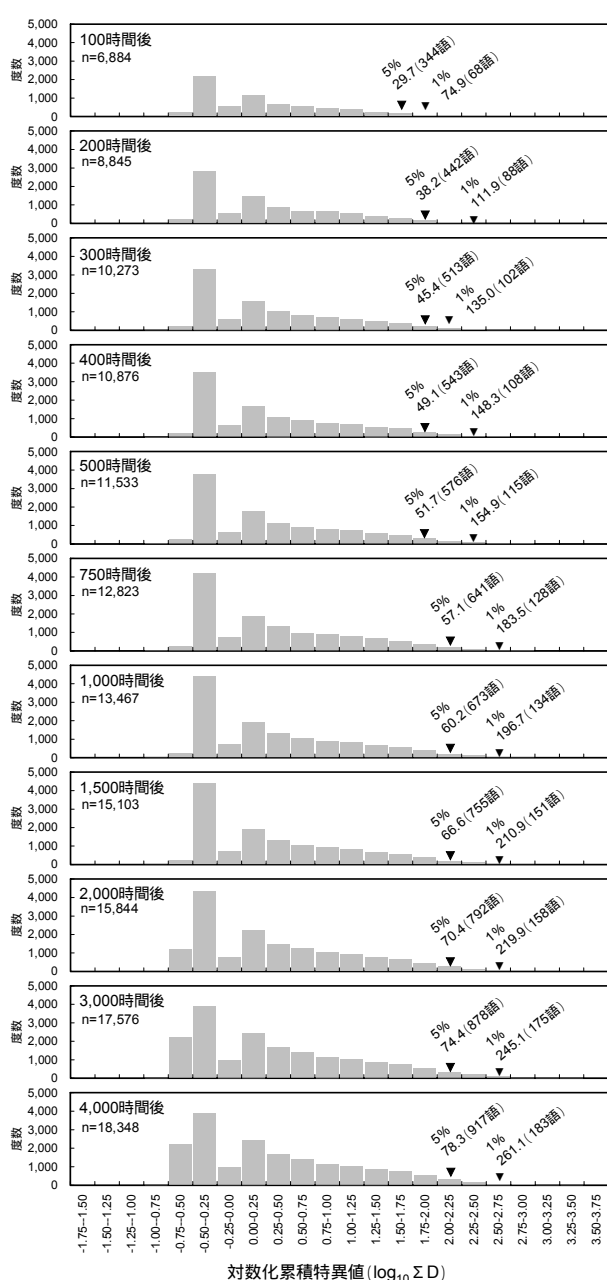


図5 累積特異値の分布特性（中越沖地震）

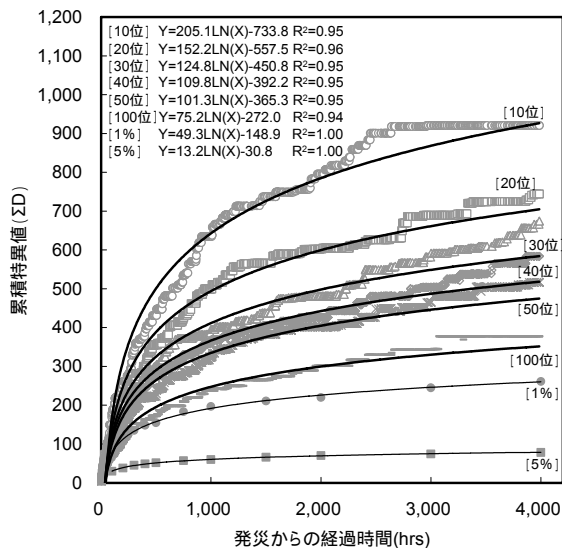


図6 中越沖地震のウェブニュースコーパスにおけるカットオフ値（累積特異値の高い上位1%，5%，10位，20位，30位，40位，50位，100位）

(2) 株価動向分析手法の援用

株取引分野では、株価チャートという株価の変動が集約・可視化されたグラフにもとづき、売買のタイミングを検討することが一般的である。株価チャートは、1) 移動平均線、2) ローソク足、3) 出来高の3つの要素から構成されている¹²⁾。1) 移動平均線は、例えば過去5日間の株価を平均した値を折れ線で結んだものであり、株価の動きを平均にならして、株価のトレンドを表す。2) ローソク足は、1日（日足）や1週間（週足）における株価データの中から、始値、終値、高値（最高値）、安値（最低値）を取り出し、前2つの値を四角形（胴体）で、後2つの値を線（ヒゲ）で表す図である。四角形の上端と下端が、始値か終値のいずれかを表し、終値と始値を比較し、終値が高ければ白い四角形、低ければ黒い四角形で表される。胴体の上下端から高値と安値の値を結んだものがヒゲとなる。3) 出来高は、証券取引所で売買が成立した株数のことで、株価チャート中の下部に棒グラフで表されるものである。これらの株価チャートの3つの要素が示す傾向には、いくつかのパターンがあり、このパターンにもとづいて株価動向の予測がなされている。

TRの手法で算出される特異値のふるまいを株価チャートを用いてグラフ化を試みたい。1) 移動平均線は、当該ポイントの前後の値を用いて移動平均値を求めるという一般的な移動平均法と異なるものの、特異値を用いて過去数日間の平均値を計算することはむずかしくない。2) ローソク足についても、1日（日足）や1週間（週足）を周期として、その中で特異値の最初の値（始値）、最終の値（終値）、および最大値（高値）と最小値（安値）を取り出すことも技術的に不可能ではない。特異値の算出（キーワード抽出）を1時間刻みで行った（5章冒頭）のは、日単位のローソク足を描画するために、日より短い時間間隔の時系列データを必要としたためである。3) 出来高については、分析対象のキーワードの単純な出現頻度（1日間や1週間におけるTFの合計）で代用できるものと考えられる。出来高は、取引された株数であり、ある銘柄の人気の度合いを表してことから、相場のエネルギーを表すものだと言われている¹²⁾。TRの

計算過程の中では、特異値や時間増加型TFIDFを計算する元になっている言葉の出現頻度が出来高に相当するものと考えられる。

TRの手法によって得られる特異値の変動傾向を株価チャートで可視化することによって、関連する災害事象の将来展開予測ができる可能性を検討するために、(1)節に示した条件で重要キーワードの同定を行った結果をもとに「観光」に着目し、株価チャートの適用を試みる。「観光」は、発災から4,000時間の時点で累積特異値が3位に位置づけられたキーワードである。この時点で1位は「訓練」、2位は「山古志」であり、前者は防災の日（9/1）に実施された訓練や中越沖地震を教訓とした地方の訓練に関する記事、後者は中越地震の現在の復興状況を報じる記事から抽出されたキーワードであった。当該の災害の将来展開を予測することを念頭におき、中越沖地震と直接の関連がある風評被害について述べられている記事から抽出された「観光」を採用することにした。

図7に「観光」の特異値のデータセットを株価チャートに表したものを示す。株価チャートはローソク足と出来高を求める単位を1日とする日足、1週間とする週足の2種類のグラフを作成した。移動平均値を計算するさいに考慮する期間には、特に定まった値はない。ここでは、日足のグラフに、移動平均値を求める期間のうち、一般的に最も短い3日間を採用し、週足の移動平均値は、これに対応して3週間の平均値を求めた。なお、重要キーワードの選定基準として、累積特異値を用いることの妥当性を検証するために、株価チャートの下部に、累積特異値にもとづくキーワードの順位の変化も併記した。特異値がピークを迎える前に、株価チャート上に特異値の上昇を予兆する傾向が見られ、さらに、これと同時にしくはそれ以前にそのキーワードが高い累積特異値を示していれば、将来展開予測を適切に行えることになる。

「観光」は7/17に出現し、特異値がおおよそ1ヶ月間上昇していき、8/11頃にピークを迎えた、8/11以前には、ローソク足が白い胴体を示しており、特異値の上昇傾向が読みとれる。7/25には、出現頻度が著しく高まり、その後、特異値が上昇していったのが分かる。累積特異値にもとづく順位を見ると、8/2（400時間）に30位以内（21位）になっていることから、特異値が高まる頃、最大値になる10日間ほど前に「観光」が重要なキーワードになっていた。なお、それよりも前の7/28（300時間）には、「観光」は累積特異値分布の上位1%以上に位置していた。

ここで、株取引分野における買いのサイン、売りのサインを移動平均、ローソク足、出来高についてまとめる¹²⁾。買いのサインはある銘柄の株価が上昇する予兆であり、売りのサインは株価が下落する予兆であるため、TRの考え方に問題を置き換えれば、前者は重要な事象になる恐れ、後者は以降、その逆の傾向になることを意味する。株価とTRの手法における特異値は異なる指標ではあるが、相場のエネルギーを表す出来高が要因となって株価を左右するという関係は、言葉の出現頻度が特異的に高まることによって特異値が増加する関係に似ている。3つの要素に関する主要な買い（売り）のサインを以下にまとめた：

- ・移動平均線：実際の株価と移動平均値が離れた場合、株価は移動平均線に近づこうとする（株価は移動平均線に収束する）。移動平均線が株価を下回っていれば、上昇する可能性がある。
- ・ローソク足：大陽線（長い陽線）の場合、買い圧力

が強く、強い上昇エネルギーがある。窓あけ急騰（連続してローソク足が並ばず、間をあけて動くもの）の場合にも強い上昇エネルギーがある。高値圏で太陽線が出現した場合には、下降トレンドに転じることが多い。

- ・ 出来高：安値圏で高い出来高が出た場合は、上昇トレンドになる傾向がある。

「観光」の特異値は、上記 3 つのパターンに合致した傾向を示した。7/17 に「観光」が出現してからは、移動平均線は、特異値の実測値の一部をプロットしているローソク足よりも下方に表れている。株の場合、株価は移動平均線に近づこうとするが、「観光」の特異値も変化の傾きが徐々にゆるやかになっている。7/25 には太陽線、7/26 には窓あけ急騰が見られたほか、7/25 には出現頻度（出来高に相当）が著しく高い値を示し、その後は特異

値の上昇が続いた。週足のチャートについても同様の傾向が読みとれる。

7/25 の記事は「新潟県を放射能の風評被害直撃、宿泊取り消し 4 万 8 千件(12:04 毎日新聞)」「闘牛 2 頭が“海水浴”新潟の海の安全の PR(19:26 毎日新聞)」という見出しであり、風評被害が顕在化しはじめ、対応の動きが見られた時期であることが分かる。特異値がピークを迎える頃は「風評被害 500 億円 原発不信、観光にも深刻な打撃(8/8 12:01 毎日新聞)」「新潟市、風評被害 2 億 2,600 万円 観光シーズン大打撃(8/9 12:01 毎日新聞)」という記事が表れ、被害が拡大していった様子が分かる。

「観光」の特異値のふるまいには、株取引における買いのサインと同じような傾向が見られたことから、株価チャートの指標や考え方を援用することによって、TR の手法で抽出されたキーワードのふるまいを予測できる可

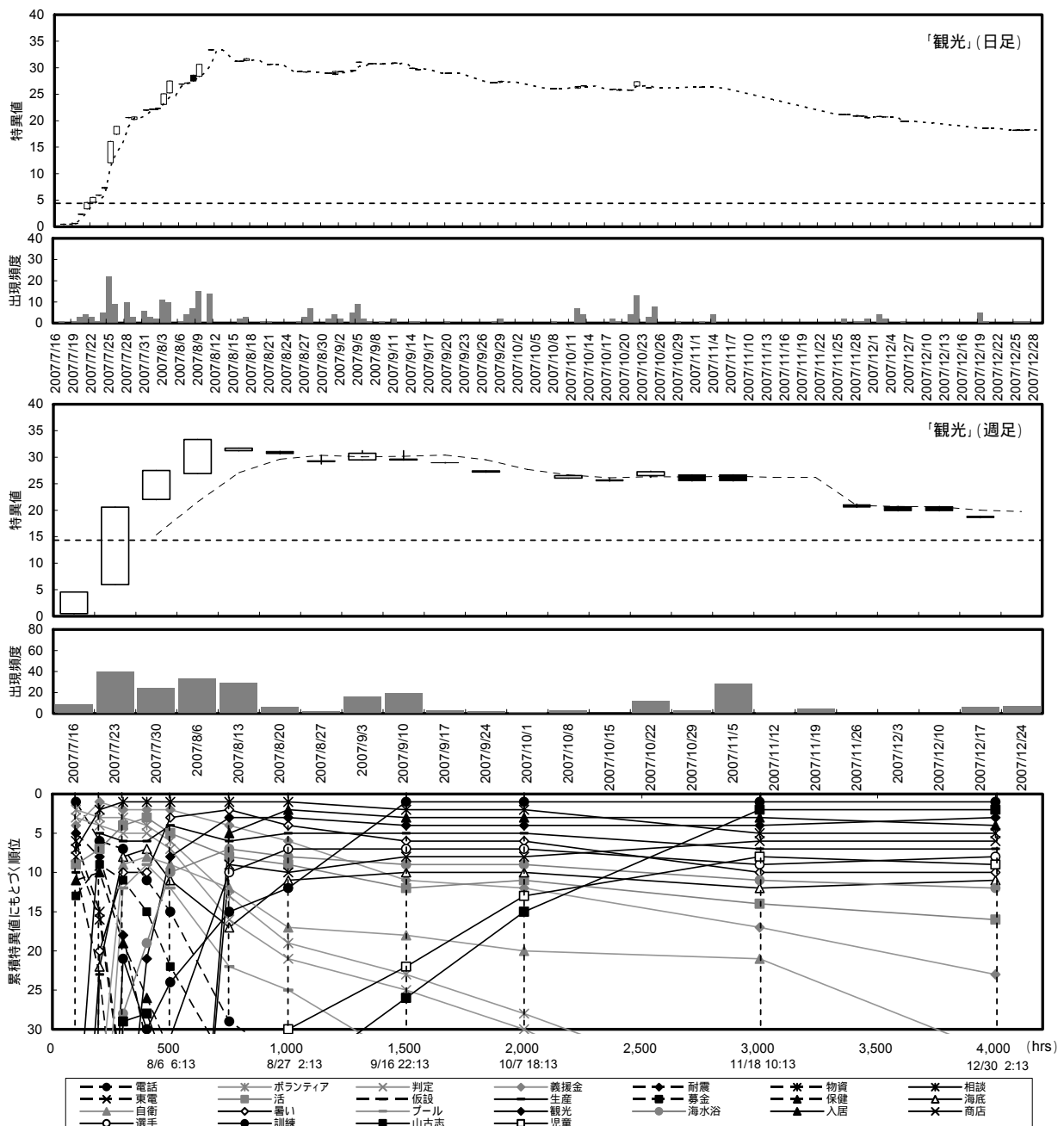


図7 「観光」の特異値のふるまいを株価チャートの指標と描画方法で図示した結果および累積特異値の順位変化

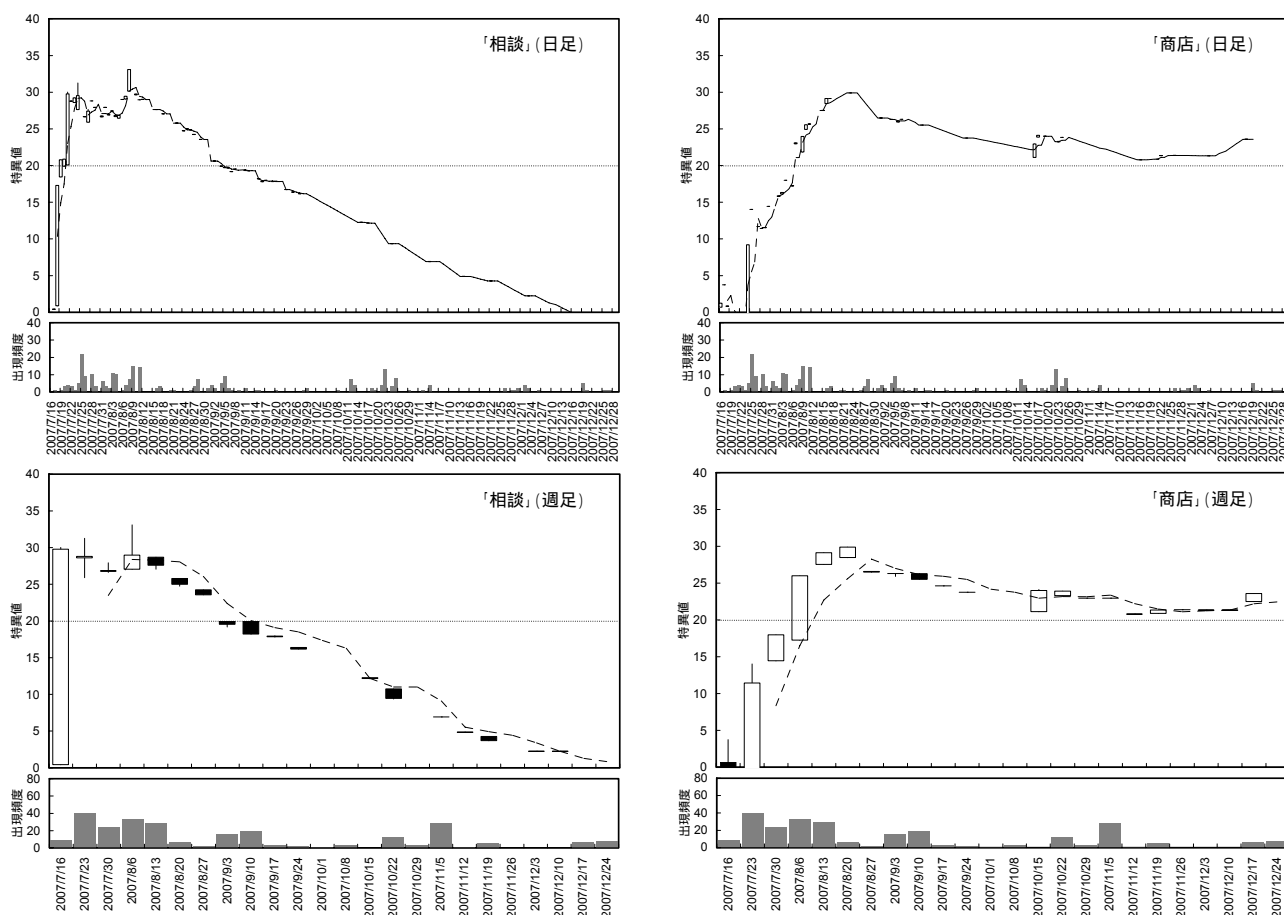


図8 特異値チャートの例（「相談」と「商店」）

能性が確認された．上記のように任意のキーワードの特異値で作成されたチャートを特異値チャートと呼ぶことにする．

6．災害・危機事象の将来展開予測手法の流れとその適用例

5章で，株価動向予測手法を用いた実験を踏まえれば，災害・危機事象の将来展開予測手法の流れは，以下のようになりまとめられる：

- 1) 災害・危機に関するコーパスを構築する：この研究で例として用いた報道資料のほか，本部会議録や住民からの要望などの言語データからなる電子コーパスを作成する．
- 2) TR を用いてキーワードを自動抽出する：キーワードを抽出するために時間刻みを 1 時間 ($\Delta t=1$ 時間) とし，災害・危機事象の将来展開予測の元になる特異値の時系列データセットを準備する．
- 3) 重要キーワードを同定する：対応に当たっている今現在までの累積特異値を求め，順位の基準や 1%，5%の基準などを目安に今後重要になると思われる事象に関連するキーワードを選択する．
- 4) 特異値チャートを描画する：日足または週足の特異値チャートを作成し，特異値が上昇・下降するパターンと照らし合わせ，今後のふるまいを予測する．

以上の予測の流れにもとづいて，「観光」以外のキー

ワードの特異値に将来展開予測手法を適用した例を示す．

4,000 時間までの時点では，「観光」の次に「入居」「相談」「商店」が高い累積特異値を示した．これは，上位 10 位に入ったキーワードである．ここでは，中越地震と特異値のふるまいが同じような傾向を示した「入居」を除く，「相談」と「商店」の特異値チャートを図示した（図8）．

「相談」の日足を見ると，7/17 に著しく長い陽線（大陽線）が見られ，その後も特異値が上昇している．その後しばらく特異値が頭打ちになり，8/8 頃に再びやや長い陽線が出現したが，その後，特異値は減少していった．これは，5章（2）節で示したローソク足から予想される株価動向パターンの一つである．これ以降に出現頻度（出来高）が高まることはなく，特異値の減少傾向とその予兆の妥当性を高める結果になっている．なお，8/8 には「柏崎市の避難所に常駐している看護師や，市健康管理センターの助産師のもとには連日，赤ちゃんを持つ母親らが相談に訪れている(7:08 読売新聞)」という記事や借家に関するトラブルやローンの問題に関する無料法律相談の窓口案内に関する記事(12:01 毎日新聞)などがあった．

「商店」の日足を見ると，7/24 に高い出現頻度と長い陽線が表れてから，特異値が上昇傾向を示している．しかし，「観光」や「相談」に見られた陽線が連なる様子は見られない．ここで，「商店」の週足をみると，日足で見られなかった陽線が連続して出現しており，上昇トレンドを予測できる可能性が見られた．キーワードが出現する時間の間隔に応じて，日足か週足かを適切に選択

することでキーワードのふるまいを予測できることが明らかになった。なお、「商店」のローソク足が長めの陽線になった 8/6 の週には、「再開 26% 新潟県、被害調査へ 柏崎・えんま通り商店街(8/7 7:51 産経新聞)」「復興対策など、3 分科会設置へ(本文:設置されるのは風評被害対策 被災地商店街の復興対策 被災者再建の 3 分科会(8/10 13:01 毎日新聞))」などの記事があった。

7. おわりに

本研究では、先行研究で開発した災害・危機に関するコーパスからキーワードを自動抽出する TRENDREADER(TR)を応用して、災害・危機事象の将来展開を予測する手法として、キーワードの重要性を評価する指標の時系列的なふるまいを予測する手法の開発を試みた。この研究の成果は以下のようにまとめられる：

- 1) ハザードや地域性が類似する 2 つの災害・危機に関するコーパスから抽出された同一のキーワードの特異値は、必ずしも同じようなふるまいを示すとは限らないことが明らかになった。
- 2) 上記のようなアプローチにもとづく手法は、災害の共通性や普遍性、災害ごとの特徴的な問題や教訓そのもの、および時系列的な展開の速さの差異などを客観的に可視化する有効な方法であることが示された。
- 3) 株取引分野で投資家やトレーダーが株価動向を予測するさいに用いる株価チャートを援用し、キーワードの特異値のデータセットを適用したところ、移動平均線、ローソク足、出現頻度(出来高)の示すパターンから、特異値が上昇するトレンドを捉えることができる例があることを確認した。
- 4) 株価チャートの指標の描画方法を援用した特異値チャートで災害・危機に関する事象の将来展開を予測する手法を提案した。

本研究では、災害・危機に関する状況を記録したウェブ報道を例にして手法の開発を行った。災害対策本部などにおける危機対応時に情報作戦を担う部門には、本部会議における議事録や住民からの声(クレーム情報)などが寄せられる。このような加工されていない生の言語資料を用いることによって、社会現象としての災害・危機事象について、より確実に効用の高い将来展開の予測が行える可能性がある。今後は、災害・危機について記録する内部情報への適用を試行し、将来展開予測の可能性とその結果の有用性について検証したい。また、情報源の異なる言語データや、言語データ以外の時系列情報にも着目し、これらのデータを重ね合わせることで、現象や人々の関心の移り変わりについて共通する部分や差異を明らかにすることで、災害過程の理解の深化を図りたい。

謝辞

本研究は、独立行政法人科学技術振興機構 独創的シーズ展開事業 大学発ベンチャー創出推進「時系列的に増加する Web 情報を対象にしたキーワード自動検出ソフトウェアの開発」(開発代表者:林春男 京都大学)および文部科学省 首都直下地震防災・減災特別プロジェクト「3. 広域的危機管理・減災体制の構築に関する研究」(研究代表者:林春男 京都大学)によるものである。

ウェブニュースの収集については、(有)アールツー・メディアソリューションおよび所属研究室大学院生のみなさんに、解析については(株)サイエンスクラフトにご助力いただきました。

参考文献

- 1) 林春男:「日本社会に適した危機管理システム基盤構築」が目指したもの、文部科学省科学技術振興調整費研究成果発表ワークショップアブストラクト集, pp. 9-23, 2006.
- 2) 佐藤翔輔, 林春男, 牧紀男, 井ノ口宗成: TFIDF/TF 指標を用いた危機管理分野における言語資料体からのキーワード自動検出手法の開発 - 2004 年新潟県中越地震を取り上げたウェブニュースへの適用事例 -, 地域安全学会論文集, No. 8, pp. 367-376., 2006.
- 3) 佐藤翔輔, 林春男, 牧紀男, 井ノ口宗成: TFIDF を用いた災害・危機に関する言語資料体からのキーワード自動検出手法の外的妥当性の検証, 地域安全学会論文集, No. 9, pp. 65-74, 2007.
- 4) A. Hartmann: Words Create Worlds, Social Work, 36, pp. 275-276, 1991.
- 5) G. Salton, C. Buckley: Term-weighting approaches in automatic retrieval, Information Processing & Management, 24, pp. 513-523, 1988.
- 6) Yiming Yang, Tom Pierce, Jaime Carbonell: A Study on Retrospective and On-Line Event Detection, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 28-36, 1998.
- 7) 松本裕治: 形態素解析システム「茶筌」, 情報処理, Vol.41, No. 11(2000), pp. 1208-1214.
- 8) Yahoo! Japan News: <http://headlines.yahoo.co.jp/hl>
- 9) 青野文江, 田中聡, 林春男, 重川希志依, 宮野道雄: 阪神・淡路大震災における被災者の対応行動に関する研究 - 西宮市を事例として -, 地域安全学会論文報告集, No. 8, pp. 36-39, 1998.
- 10) 田中聡, 林春男, 重川希志依: 被災者の対応行動にもとづく災害過程の時系列展開に関する考察, 自然災害科学, Vol.18, No. 1, pp. 21-29, 1999.
- 11) 林春男: 率先市民主義 防災ボランティア論 講義ノート, 晃洋書房, pp. 52, 2001.
- 12) 渋谷高雄: 世界一わかりやすい 株価チャート実践帳, あさ出版, 191pp., 2008.

(原稿受付 2008.5.24)

(登載決定 2008.9.13)