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年新潟県中越地震災害を取り上げたウェブニュースへの適用事例－

The Development of an Algorithm Using the TFIDF / TF Index to Extract Automatically the Set of Keywords of Corpus about Fields Related to Emergency Management
-A Case Study Utilizing Web News Articles for the 2004 Niigata-Ken-Chuetsu Earthquake Disaster-

佐藤 翔輔[1]，林 春男[2]，牧 紀男[2]，井ノ口 宗成[1]

Shosuke SATO[1], Haruo HAYASHI[2], Norio MAKI[2], and Munenari INOGUCHI[1]

[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

Based on the web news archive of a total of 2,623 articles highlighting the 2004 Niigata-ken-Chuetsu Earthquake disaster, we developed an algorithm to extract automatically the set of keywords which can characterize any time phase of disaster process from the start of an event to the termination of Emergency Operation Center. And we applied a visualization method with use of extracted keywords set and their feature quantify. Extracted keywords by the algorithm and the text mining result helps Cross Media Database (XMDB) users to search information and intelligence from corpus about fields related to emergency management.

Key Words : *keyword extraction, automatic algorithm, database management, corpus, text mining, emergency internet news, the 2004 niigata-ken-chuetsu earthqyake disaster*

１．はじめに

（1）危機管理分野に関する情報共有のためのクロスメディアデータベースの必要性

著者らは，防災研究者や防災実務者の間で情報を共有・交換するための検索・表示機能を含む包括的なデータベース（Cross Media Database 以下，XMDB）の開発を試みてきている[1)]．

防災の世界は，多くの学問分野の協働を必要とする学問領域であるとともに，実務者と研究者の協働を必要とする実学的な分野である．これは，防災を取り巻く世界全体に精通することは困難であることを意味している．個々の分野に対する知識の不足によって理解が妨げられるだけでなく，学問分野ごとの手法で情報が収集，蓄積，集約されており，それぞれの領域に合ったフォーマットをもつデータや研究成果はしばしば使いづらく，理解しがたいものになっている．そのため，防災の世界では，学問分野を異にする研究者の間，また，防災の実務者と研究者のコミュニケイションも困難なものになっている．

このような背景から，防災の世界において実務者や研究者の容易な情報交換を可能にし，横断的な研究の推進や研究成果の実務領域への浸透を図ることを目標として，他の分野の研究者や実務者にも利用されるべき自分野の防災に関連したデータや情報，研究成果を媒体の種類による制約を受けずに，ユーザーが親しみやすいインターフェイスを使って，いつでもどこからでも，情報の検索を可能にするような研究支援や実務支援の基盤構築の必要性が高まっている．

（2）社会現象としての災害の記録のデータベース化の必要性と問題点

以上に述べたXMDBに蓄積すべきデータや情報は，強震計によるゆれの観測結果や気象庁が観測する全国の降雨量などの自然現象に関するデータや情報に限らない．研究の発展や，研究成果と過去の教訓の実務分野への浸透を図るためには，体験談記録，災害対応の記録（様式やメモ），被害報告，刊行資料，新聞記事やウェブニュース記事などの社会現象としての災害に関するデータや情報もデータベース化の対象になる．

防災の世界において，災害に関する社会科学的な研究への取り組みが盛んになって久しい[2)]．災害の研究は，自然現象としての災害を対象とする力学を応用した自然科学的な研究に加え，災害を体験する被災者・災害対応従事者・被災地外の人々を含む社会，災害からの復興問

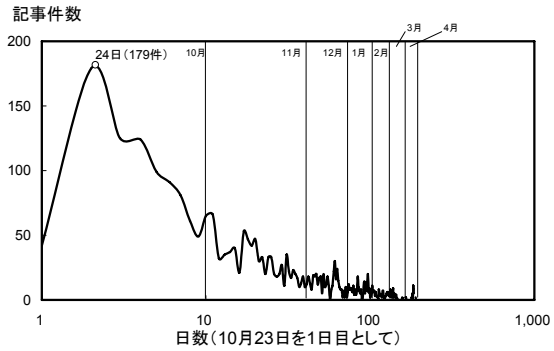

図 1　新潟県中越地震災害に関するウェブニュースの発信状況

題を扱う社会現象としての側面を考慮した研究が，1995年の阪神・淡路大震災や2001年での米国テロ事件の発生を契機にして数多く取り組まれている．社会現象を取り扱う研究も，自然科学の枠組みと同様に災害状況の記録のデータベース化が要になっている．自然災害科学の領域では，強震計によるゆれの観測結果，気象衛星による雲の動きの観測結果などをもとに様々な解析をおこない，地震や豪雨という自然のハザードの発生過程に対する理解の深化が図られたり，シミュレイションの入力外力として用いられ，構造物の耐力向上に資する研究がなされている．社会現象を扱う領域においても，自然現象の理解や構造物の耐力向上に向けた自然災害科学の研究方法と同じように，データや資料のデータベース化をおこない教訓や知識を抽出・体系化し，効果的な災害対応を実現する材料を準備することが求められる．また，研究のみならず，過去の災害対応に関する種々の記録は，実務者が目を通すべき重要な情報資料として位置づけられる．

ところが，社会現象に関する災害下における社会現象の記録は，データの形態が言語資料（テキスト資料）であるために，XMDBへの蓄積や情報検索のさいには，以下のような問題が発生する．

まず1点目として，データベースへの蓄積のさい，単位データ（1つのレコード）の内容を表すキーワードの付与には，多くの人的資源と専門知識を要することが挙げられる．XMDBは，時間・空間・テーマにもとづく情報検索の機能を搭載しているため，蓄積されるデータには，データの作成日時などの時間情報，データがもつ位置情報，データの内容を代表するキーワードという3種類のメタデータをレコードに付与することが必須となっている．なお，以上のようなメタデータを付与することは，諜報活動の場面においても重要な手続きとして位置づけられており，情報資料を管理する上で，またトレンドを分析する上でも欠かすことのできない手続きとなっている[3]．このデータの内容を代表するキーワードを付与する作業には，防災分野に対する包括的な理解をもった人的資源が必要になる．しかし，現実にそのような人物は存在せず，災害の発生を契機として，様々な情報源から発信される膨大な量のデータを人が一つ一つ判読し，キーワードを付与することは実質不可能であるのみならず，ここには作業者の恣意性（主観的感覚）が介入してしまう．

2点目の問題は，どのようなキーワードを用いて情報検索を行えばよいのかという点である．防災の世界に対して包括的な理解をもった人や，個々の災害の事例に詳しい人であれば，既存の知識をもとに情報検索に要するキーワードを容易に想像することができる．しかし，専門知識をもたない実務者が適切な検索キーワードを想像することは難しいことは当然のこと，研究者自身もそれぞれの研究分野に偏ったテーマに対する知識しかもっておらず，災害事例のすべてに精通しているわけではない．

（3）研究の目的

これまで述べてきた社会現象としての災害を記録する情報資料（主に言語資料）のデータベース化の必要性と，データベース化や情報検索に伴う問題を受けて，本研究では，発信・収集された言語資料から，メタデータとして付与すべき情報内容を表すキーワードを自動的・客観的に生成する手法を構築することを目的とする．これは，収集された言語情報資料をXMDBに蓄積するさいに，キーワードを付与する手続きにおける前述した問題の解決を目指すものである．また，キーワードの生成・検出結果を用いて，キーワードの特徴の度合いを定量的に評価し，言語資料体のもつ情報を縮約してユーザーに提示するテキストマイニング手法の構築を試みる．ここに構築するテキストマイニング手法は，データベースに蓄積される大規模な量のテキストで構成される言語資料体がもつ情報量を縮約したかたちでユーザーに提示することにより，容易な情報検索を促すことを意図している．

なお，この研究の後半では，開発した手法を災害事例の記録に適用し，その結果を用いて，災害過程に関する知見に関する考察を試み，キーワード自動検出手法とテキストマイニング手法の有効性を検討する．

２. 検出手法の適用対象の選定と収集

この研究では，キーワードを自動的に検出する手法の検討や構築，適用を試みるための言語資料体として，危機についてウェブ上に配信されたウェブニュースを試験的に用いることにする．災害を社会現象としての観点から記述する言語資料には，ウェブニュースのほかに，体験談記録，災害対応の記録，被害報告，刊行資料，新聞記事などの紙媒体に書かれた印字や手書き文字で記録されたものがある．これらは，OCR（Optical Character Reader）の技術によってテキストをデジタル形式に変換し，テキスト情報を加工してキーワードの生成をおこなうことが可能であると考えられる．それに対してウェブニュースは，デジタル媒体で発信されており，そのまま蓄積や加工が容易であること，発災からの災害の情報を継続的に記録しつづけていることから，キーワードの検出手法の適用実験に用いやすいものと考えられた．また，ウェブニュースは，ネットワークメディアであるために，全国で発生した災害についての収集（観測），データベース化が可能である．以上のような点から，この研究では，危機・災害に関するウェブニュースをキーワード検出手法の適用対象とし，開発を進めていく．

この研究では，2004年新潟県中越地震（平成16年10月23日17:56発生，M 6.8）について発行されたウェブニュースを用いる．新潟県中越地震災害を対象としたのは，インターネットの普及以降，我が国で発生した災害の中でも比較的規模の大きな災害であり，多くのニュース記事を収集・分析できると考えたためである．

平成16年（2004年）10月23日以降に代表的なポータルサイト[4]のニュースコンテンツ上に発信された新潟県中越地震災害に関連するニュースを収集し，発信日時，発信新聞社，タイトル（見出し），記事本文，をフィールドにしてデータベースを作成した．すべての記事に対して，ポータルサイト上に更新されてから24時間以内に収

集する作業をおこなった．収集した期間は発災からは翌年4月30日までのおよそ6ヶ月間である．収集したウェブニュースは2,623件である．**図1**に発災日からの経過日数を対数目盛りでとり，毎日の記事収集件数を示した．地震が発生した当日は，18時59分に最初のニュース記事がアップデイトされ，当日中には42件発信された．記事件数が最も多かったのは地震が発生した翌日の24日で179件だった．

３．キーワード候補の同定

この章では，ウェブニュース記事から特徴的なキーワードを自動検出する手法の開発するために，キーワードの単位と不要語の除去方法についての検討をおこない，手法の適用を試みる．**（1）節**では，キーワードとして採用すべき言葉の単位を検討し，**（2）節**では，**（1）節**で決定した言葉の単位の中でも，キーワードとして適切ではないものを取り除く方法について検討する．

（1）キーワードの候補となる単位の決定

ここでは，キーワードとなる言葉の単位を決定し，実際にキーワードの候補になる単位にする解析を試みる．

日本語は，段落，文，文節，単語，文字などの単位に分割することができるが，キーワードとして一般に用いられる単位は単語である．しかし，国語学上，単語に対する厳密な定義はない．たとえば，「新潟県中越地震」であれば，これをそのまま単語として捉えることもできるが，「①新潟 / 県 / 中越 / 地震」，「②新潟県 / 中越 / 地震」，「③新潟県中越 / 地震」などのように分割することができ，考え方や視点によって，そのパターンは複数存在するため，このような複合語について配慮することは客観的に単語を同定することを困難にする．

キーワードとしての単語を切り出す解析法として，一般に形態素解析が用いられてる．形態素解析とは，自然言語で書かれた文を形態素と呼ばれる文法的に意味をもつ最小単位の文字列に分割し，品詞情報を付加する自然言語処理の基礎技術である． 形態素解析の結果の例を示す：「新潟 / 県 / 中越 / 地震 / は / 住民 / の / ライフライン./ に / も / 甚大 / な / 被害 / を / 及ぼし（及ぼす） / た / ．」．上述した例の①のような解析結果が出力されるほか，「及ぼし（及ぼす）」のように，活用形をとった形態素に対しては基本形をも出力する．この形態素解析は，現在の技術水準でおおよそ96～98%以上の精度を達成している[5]．

ここでは，形態素の単位をキーワードの単位として採用することにする．形態素の単位では，「新潟県中越地震」のような複合語を捉えることはできない．しかし，現段階では単語という適切な概念や定義は存在せず，また言語データから切り出す解析法も存在しない．形態素の単位であれば，高い精度での解析が可能であることから，この研究では形態素の単位をキーワードの候補とする．

ウェブニュース全記事に対して，形態素解析の結果を試みた結果，15,211種類の形態素（合計623,765の形態素）が得られた．

（2）不要語の除去

形態素解析によって得られる形態素群の中には，キーワードとして適さないものが存在する．ここにいうキーワードとして適さない語とは，助詞の「が」や「を」のように，主にそれ自体に意味をもたないもの形態素のことを指す．一般に，このような言葉を不要語と呼ぶ．不要語のような言葉自体からは，意味や内容を捉えることはできない．

本節では，以上のような不要語のもつ問題点から形態素解析によって得られる各形態素の品詞に着目して，キーワードとして適さない形態素を除去することを検討する．以下，本研究で用いた形態素解析システム[6]のもつ品詞体系が採用している品詞情報にもとづいて，不要語とする品詞を決定していく．

助詞（「が」，「を」），助動詞（「れる」，「られる」），接続詞（「しかし」），記号（「句読点」）は，文法的な役割をもつ品詞で，内容的な意味をもたない品詞であり，キーワードとしては適さない．また，他の形態素と結びつくことで意味をなす品詞は，一つの形態素では意味を捉えることはできないためキーワードとして適さない．これには，名詞・動詞・形容詞のうち，非自立や接尾のかたちをとるもの（「こと」，「しまう」，「らしい」），接続詞的な名詞（「対」，「兼」），接頭詞（「お」，「約」），連体詞（「この」，「その」）が該当する．そのほか，他の語を指すためにそれ自身では意味を捉えることができない代名詞（「それ」，「わたし」），話の間をとるためだけ用いられるフィラー（「ええと」，「うんと」）もキーワードとして適さない．また，あいさつやあいづちなどの感動詞（「おはよう」，「いいえ」）は主に会話の中で用いられることから，災害事象との関係は薄いものと考えられる．

以上の品詞を取り除けば，名詞，動詞，形容詞のうち，非自立や接尾のかたちを取らないものと副詞がキーワードの候補として採用されることになる．

品詞情報をもとに不要語を除去した結果，**（1）節**で求められた15,211種類の形態素は，14,109種類にまで減少した（のべ521,240の形態素）．14,109種類のうち，地震の発生から1～10時間で1,122種類の形態素（72記事），10～100時間で3,581種類の形態素（481記事），100～1,000時間で5,691種類の形態素（1,230記事），1,000～4,529時間で2,716種類の形態素（840記事）が出現した．

４．特徴量にもとづくキーワードの検出

本章では，ニュース記事から抽出したキーワードの候補に重みを与えることよって，キーワードがどれだけ特徴的であるのか，ある時間の変化を代表するキーワードとしてどれだけ重要なのかの評価を試みる．

（1）キーワードの重み付け

ある時点でのキーワードに，特徴の度合いを表す指標の情報が付加されていれば，指標の評価結果にもとづき，より特徴的なキーワードを同定することができる．本節では，キーワードに特徴の度合いを表す指標を与えることを検討する．

ある時点で，ある事柄がウェブニュース上で中心的に発信されている場合，ある事柄の意味を表す言葉は多く出現する可能性がある．しかし，頻出するキーワードの中には，1）どのようなニュース記事であっても，文書を構成する上で多用されるキーワード，2）一部のニュース記事の中で頻出しているキーワードの2種類があることが想像される．ニュース記事を特徴的に表すキーワードとは後者を指す．

後者のようなキーワードに対して高い重みを与える指

標として*TFIDF*がある．*TFIDF*指標は以下の式で算出される：

$$TFIDF(t_i, d_j) = TF(t_i, d_j) \cdot IDF(t_i) \quad [1]$$

$$IDF(t_i) = \log_{10} \frac{N}{DF(t_i)} \quad [2]$$

t_i：*TFIDF*(t_i, d_j)を算出する対象となるキーワード．*i*はその識別子を表す添え字．

d_j：*TFIDF*(t_i, d_j)を算出するキーワードt_iが含まれている文書．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，記事を文書単位とする．*j*は識別子を表す添え字．

N：文書（記事）の総数．

TF(t_i, d_j)：Term Frequencyの略．キーワードt_iが記事d_jに出現した回数．

IDF(t_i)：Inversed Document Frequencyの略．

DF(t_i)：Document Frequencyの略．キーワードt_iが出現する文書数．

以下，特に断りのない場合，各指標中のかっこ（(t_i, d_j), (t_i)）を省略して記述する．*IDF*は，全文書数に対するキーワードt_iが出現した文書数の比の逆数である．つまり，どの記事にも現れるような形態素については低い重みを，他の記事にあまり現れないような形態素には高い重みを与えることになる．これと*TF*との積をとった*TFIDF*は，記事の中にいかに多く出現し，いかに他の記事に出現していないかを表す指標であり，キーワードの特徴の度合いを評価している指標と言える．

この研究では，ある記事d_jに対する*TFIDF*を求める場合，最終的に収集された全2,623件の記事にもとづく*N*や*DF*を用いることはせず，記事d_jが発行されるまでの時間に発信されていた記事の数にもとづく時間を考慮したN_j（記事d_jが発信された時点までの記事の総数）や，*DF*(t_i, d_j)（記事d_jが発信された時点までの形態素t_iの出現文書数）を用いて，記事d_jが発信された時点で逐次*TFIDF*を計算することにする．危機管理分野における言語資料は，危機や災害の発生から時間の経過に伴って，言語資料の数が増大していく．通常の*TFIDF*は*N*と*DF*が一定であり，時系列的に増加する言語資料から抽出された形態素に対する重み付けには対応していない．本研究では，全文書数と任意の形態素が出現する文書数を時間情報にもとづいて変化するパラメータとし，*TFIDF*を修正して用いることにした．なお，このようにして*TFIDF*を求めた場合，記事d_jが発行された時点で，はじめて出現した形態素の*TFIDF*を評価すれば．*DF*は1となり，*IDF*は高く評価されることとなり，初出の形態素に高い重みを与えることになる．この時間の概念を考慮した指標を修正*TFIDF*と呼ぶことにする．

この研究では，記事ごとに解析した形態素と*TF*と修正*TFIDF*で構成されるデータベースをコーパスと呼ぶことにする[(1)]．

（2）修正TFIDFとTFの関係を用いた特徴量の評価

単に修正*TFIDF*の値だけではキーワードが特徴的であるか否かを評価することは難しい．ある時点までの*TFIDF*の値が高く評価されるパターンには，*TF*の値が低くとも*IDF* が高い（*DF*が低い）ために*TFIDF*が高い値で求められる場合と，*IDF* が低くとも（*DF*が高くとも）*TF*が著しく大きいために*TFIDF* が高く算出される場合とがある．*TF*が著しく大きいということは，その言葉の一般性が高いために記事を記述する上で何度も用いなければならないような言葉である可能性が高い．単純に*TFIDF*の値によってキーワードが特徴的であるかどうかを単純に評価することはできない．

ある時点における情報が特徴的であるということは，前の時点までに語られているキーワード群と，ある時点で語られているキーワード群とを比較することから把握できると考えられる．両者に差が生じていれば，任意時点の前後に大きな質の違いがあったことを意味していると思われる．つまり，ある時点のコーパスと，ある時点から任意の時間が経過した分のコーパスを比較することにより，情報の質の変化を捉え，その変化をもたらしたキーワードを同定できる可能性があるものと考えられる．

以下では，ここまでに算出した修正*TFIDF*と*TF*をもとに，ある時点と次の時点のコーパスの特性の比較する手法について検討する．

図2に発災からそれぞれ10時間，100時間，1,000時間，4,500時間までの形態素ごとの*TF*の累積値と修正*TFIDF*の累積値の関係をプロットした．

*TF*の累積値と修正*TFIDF*の累積値の間には，$Y=aX^b$の関数形（累乗関数）で表される強い関係があった．Y=aX+bのような関数（線形関数）で両者の関係をみると，10時間でY=0.16X+3.14（R^2=0.24），100時間でY=0.07X+10.47（R^2=0.13），Y=0.11X+18.46（R^2=0.15），Y=0.15X+22.27（R^2=0.18）と累乗関係のものには及ばなかった．なお，ここに示した発災からの経過時間以外についても同様の傾向があり，サンプル数（キーワード数）が少ない10時間までの*TF*の累積値と修正*TFIDF*(の累積値の関係以外については，累乗関数でR^2が0.90～0.99，線形関数でR^2が0.13～0.17であり，*TF*と修正*TFIDF*の累積値の間には，累乗関数の関係が系統的に存在することが明らかになった．

図2のような関数関係は，近似曲線の近傍にあるキーワードは*TF*の累積値と修正*TFIDF*の累積値の関係が，コーパスの平均的な関係と同じような傾向にあることを意味している．このような傾向をもつキーワードは，平均的な出現パターンを呈しているものと考えられる．したがって，実際の修正*TFIDF*の累積値が，近似曲線にもとづく推定値を下回る場合，コーパスの平均像からみて修正

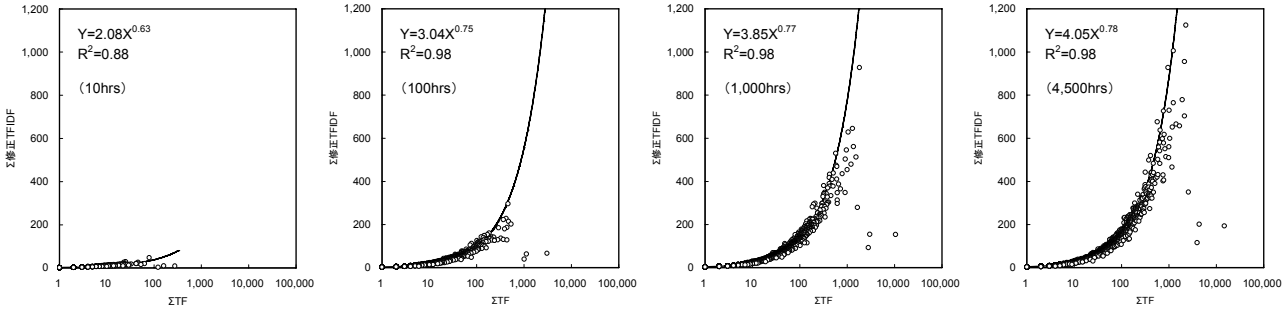

図 2　Σ修正 TFIDF と ΣTFIDF（発災から 10 時間，100 時間，1,000 時間，4,500 時間）

*TFIDF*の累積値が低い，つまりあまり特徴の度合いが高くないことを表す．逆に，実測値が推定値を上回る場合は，その逆で修正*TFIDF*が高く，特徴的なキーワードであることと言える．以上のような評価は，実際の修正*TFIDF*の累積値と，近似曲線にもとづく推定値との差（残差）を求めることによって可能になる．

以上の関係を応用し，**図3**のようなモデルで任意時点のキーワードを特徴的の度合いを評価する．

図3の左側には，ある*t*-Δ*t*から単位時間幅Δ *t*経過するさいのコーパスの変化を模式的に表した．このような関係は次式で表すことができる：

$$\overline{C_t} = C_{t-\Delta t} + C_{\Delta t} \qquad [3]$$

$\overline{C_t}$ ：ある時間 *t* におけるコーパス

$C_{t-\Delta t}$ ：ある時間よりもΔ *t* だけさかのぼったコーパス

$C_{\Delta t}$ ：ある時間 *t*-Δ *t* から *t* までに増加したコーパス

$C_{\Delta t}$ にそれまでに出現したキーワードが多く含まれていたり，出現頻度もあまり高くないような形態素のみが存在しているような場合には，**図 3** の右上側に示したように *TF* の累積値と修正 *TFIDF* の累積値の関係は，*t*-Δ *t* の時点のコーパスで構成された場合と *t* の時点のコーパスで構成された場合では大きな差は生じない．それに対して，*t*-Δ *t* までに出現しなかったようなキーワードがΔ *t* の中で出現したり，高い頻度で現れるような形態素が存在する場合には，*t* の時点でのコーパスが大きく変化し，*TF* の累積値と修正 *TFIDF* の累積値の関係を表す曲線の形状も大きく変化する．

つまり，ある時点 *t* での修正 *TFIDF* の累積値と，*t*-Δ *t* の時点でのコーパスで構成された関係式にもとづく推定値との残差が，このΔ *t* の間のコーパスの変化そのものを表し，残差が大きい形態素こそがΔ*t* 間に発生した言語資料の内容を代表するキーワードであると考えられる．

本手法ではΔ *t* での情報内容の質的な変化を表すキーワードの特徴量を評価する指標として，任意時間 *t*-Δ*t* のコーパスで構成される *TF* と修正 *TFIDF* の累積値にもとづく関係式による修正 *TFIDF* の累積値の推定値と *t* の時点での修正 *TFIDF* の累積値の実測値との差分（残差）を採用することにする．ここに残差が著しく高かったキーワードを特徴語（残差値：正），著しく低かったキーワードを一般語と呼ぶことにする（残差値：負）．

５．キーワード自動検出手法の適用結果と考察

ここまでに議論した内容から，言語資料体からキーワードを検出する方法は以下のようにまとめられる：

1) 危機の事例のテキストデータ（この研究では，ウェブニュース）のデータベースを構築する．
2) テキストを形態素に分割し，品詞情報を付加する．
3) 品詞情報にもとづき，非自立と接尾以外の名詞・動詞・副詞・形容詞を抽出する．
4) 形態素について，文書（ここではウェブニュース記事）ごとに *TF* と時間情報にもとづく修正 *TFIDF* を求める．
5) ある時点 *t*-Δ*t* から *t* の間における特徴的なテキストを代表するキーワードを抽出するため，*t*-Δ*t* までのコーパスにおける *TF* の累積値と修正 *TFIDF* の累積値の関係式を求め，それにもとづく *t* の時点での修正 *TFIDF* の累積値の推定値と実測値との差を求める．この残差値をあるΔ*t* に出現したキーワードの特徴量とする．
6) 最も大きい残差値から任意の上位数までのキーワードを選定し，当該キーワードが検出された記事にキーワードを言語資料のメタデータとする．

図 4 には，各手続きでのインプットとアウトプット，および参照すべきものとツールとともにキーワードの検出過程を示した．この図が示すとおり，本手法は人の主観的な判断を用いず， *TFIDF* 指標や残差値による定量的な指標を用いて構成されていており，連続したプロセスから成り立っているため，ツールと参照すべきものが適切に準備されていれば，過去の危機の記録をインプットとし，一連の過程を通して自動的・客観的に最終成果物であるキーワードを検出することができる．

本章では，開発したキーワード自動検出手法を，2004年新潟県中越地震災害を取り上げたウェブニュースに適用することを試みる．

青野ら[7]と田中ら[8]によって，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

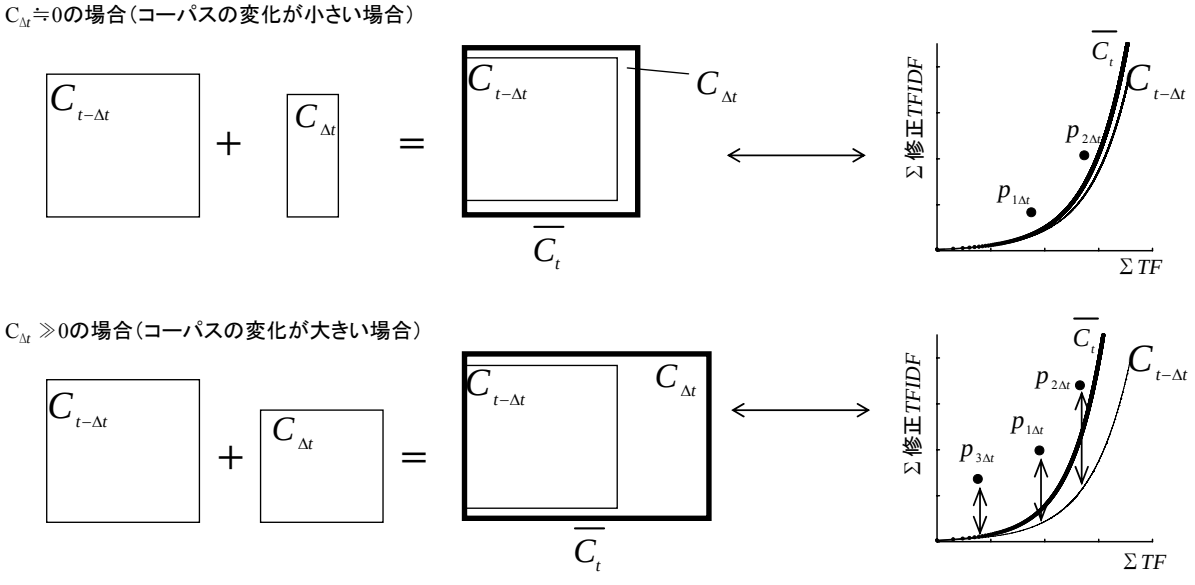

図 3　コーパスの変化とΣ修正 TFIDF とΣTF の関係

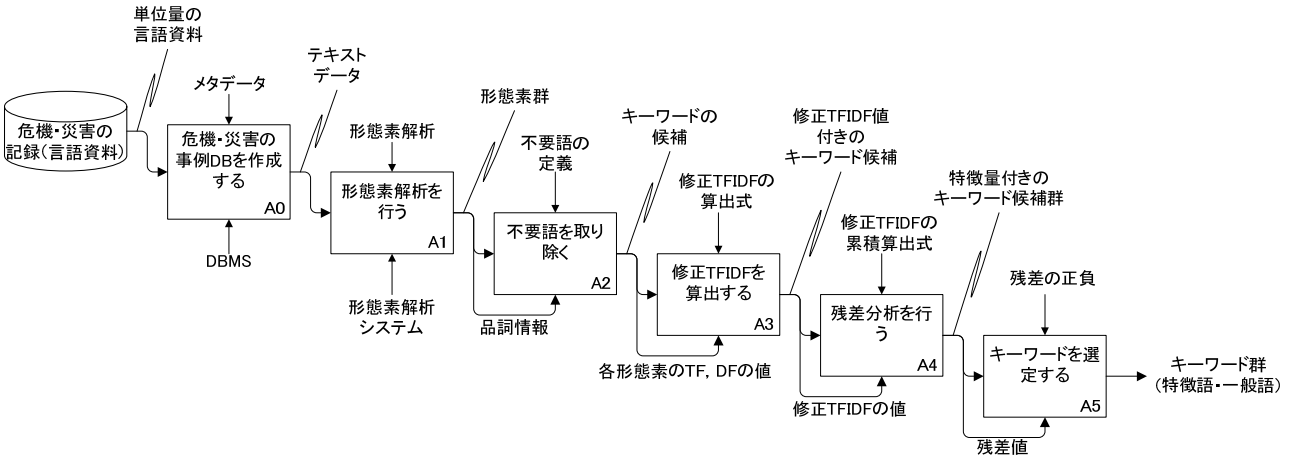

図 4　本研究で開発したキーワード自動検出アルゴリズム

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言われている．1～10 時間は失見当期と言われ，災害による大規模な環境の変化により何が起こっているのかを把握できない時期で，次の 10～100 時間は被災地社会の成立期にあたり，命を守る活動や避難所の開設などが行われる．100～1,000 時間は被災地社会が維持される時期で，社会のフローを回復し，被災者の生活を安定させる時期である．1,000 時間以降は，現実への帰還の時期に当たり，社会のストックの再建が行われる．なお，同様の災害過程のモデルは，立木 [9] によって 2001 年の米国テロ事件について報じたニューヨークタイムス関連記事を対象にした計量言語学的な解析によって外的な妥当性が確認されている．

本論文では，以上の災害過程のモデルに基準とし，1～10 時間，10～100 時間，100～1,000 時間，1,000 時間以降の 4 つの時間フェーズごとに，キーワード検出に用いる Δt をそれぞれ，1 時間，10 時間，100 時間，1,000 時間に設定してキーワード検出を試みた（**図 5**）．

図 5 に，検出されたキーワードがもつ特徴量（残差）のプロットの分布を示した．**図 5** では，時間断面ごとに検出されたキーワードの特徴量が概ね上位 3 位のものまで，および概ね下位 3 位までのものについて示した．

図 5 で検出されたキーワードにはどのようなものがあったのかをより多く観察するために，特徴量が各時間断面で上位 10 以上になったものについて，その回数を集計したものを**表 1** に示した．**表 1** には，上位 10 以上になった回数が 2 回以上のキーワードについて示してある．検出された主なキーワードの中としては，「ボランティア」が最も多く，「IC（インターチェンジ）」「断層」がつづいている．

次に，検出されたキーワードの特徴量が時間の経過ともに変化していくのかについて考察する．林 [10] によれば，災害対応には大きな 3 つの活動が存在すると言われている．第 1 は，命を守る活動で救命救助，安否確認，二次災害の防止などが挙げられる．第 2 は，社会のフローを安定させる活動で，避難所の開設，ライフラインの復旧，代替手段の提供などがこれにあたる．第 3 の活動は，社会のストックを再建させる活動で，都市・経済・生活の再建を図ろうとするものである．**図 5** および**表 1** の中からこれらの活動に関連のあるキーワードに着目し，それらの時系列的な展開についての観察を試みる．

図 6 には，命を守る活動に関連のあると思われる「電話」「死亡」「派遣」「安否」の特徴量の時間的な変化を示した．「電話」と「安否」は「地震の発生直後から、安否確認や問い合わせの電話が集中し(10/24 1:19 読売新聞)」という安否確認に関する記事などにあり，「死亡」は死者発生の報じるもの，「派遣」は「警視庁は 23 日夜、警察庁長官からの出動命令を受け、新潟県の被災地に広域緊急援助隊を派遣した(10/23 22:05 毎日新聞)」などの記事に存在している．これらのキーワードは，発災から 10～100 時間の間で特徴量のピークを迎え，それ以降，特徴量が負の値をとるようになり，一般性の高いキーワードとして位置づけられた．「死亡」については，100 時間以降で特徴量が最も低い負の値を示している．これは，「新潟県中越地震は 23 日で発生から 1 ヶ月を迎えた．死者は 40 人、重軽傷者は約 2,860 人に上り、家屋被害は約 5 万 1500 棟となった(11/23 1:25 共同通信)」のように，震災の被害の要約が何度も報じられたため，コーパス全体における「死亡」の一般性が高くなったと思われる．

図 7 には，社会のフローを回復させる活動に関連すると思われる「ボランティア」「IC」「レール」「トンネル」について特徴量の変化を示した．「ボランティア」は，社会のフローを回復させるさいの代替機能を補助する役目を担い，「IC」「レール」「トンネル」は交通系のライフラインを構成するものである．これらは，「トンネル」を除いて発災から 100～1,000 時間の間に特徴量が最大となっていた．交通系ライフラインは，「関越道は，上り線の長岡ジャンクション(JCT)―湯沢 IC 間，下り線の月夜野 IC―長岡 JCT 間で通行止めとなっている(10/26 0:27 共同通信)」のような被害についての報道ともに，「関越自動車道上下線の長岡ジャンクション－長岡 IC 間，上りの六日町 IC－湯沢 IC 間の規制も解除した(10/27 1:58 共同通信)」のように復旧の様子についての情報もこの間に発信されている．「レール」「トンネル」は新潟県中越地震のさいに発生した新幹線脱線事故について「JR 東日本は二十六日，脱線した上越新幹線「とき 325 号」をレールに戻す作業を二十七日から開始すると発表した(10/27 2:28 産経新聞)」のような復旧への動きが報じられていた．以降も「トンネル」については，何度も記事中に出現し，1,000 時間以降で特徴量は負の値をとることになる．

最後に社会のストックを再建する活動について同様の分析を試みる．**図 8** には，「入居」「判定」「補助」「移転(集団移転)」の特徴量の時間的な変化について示

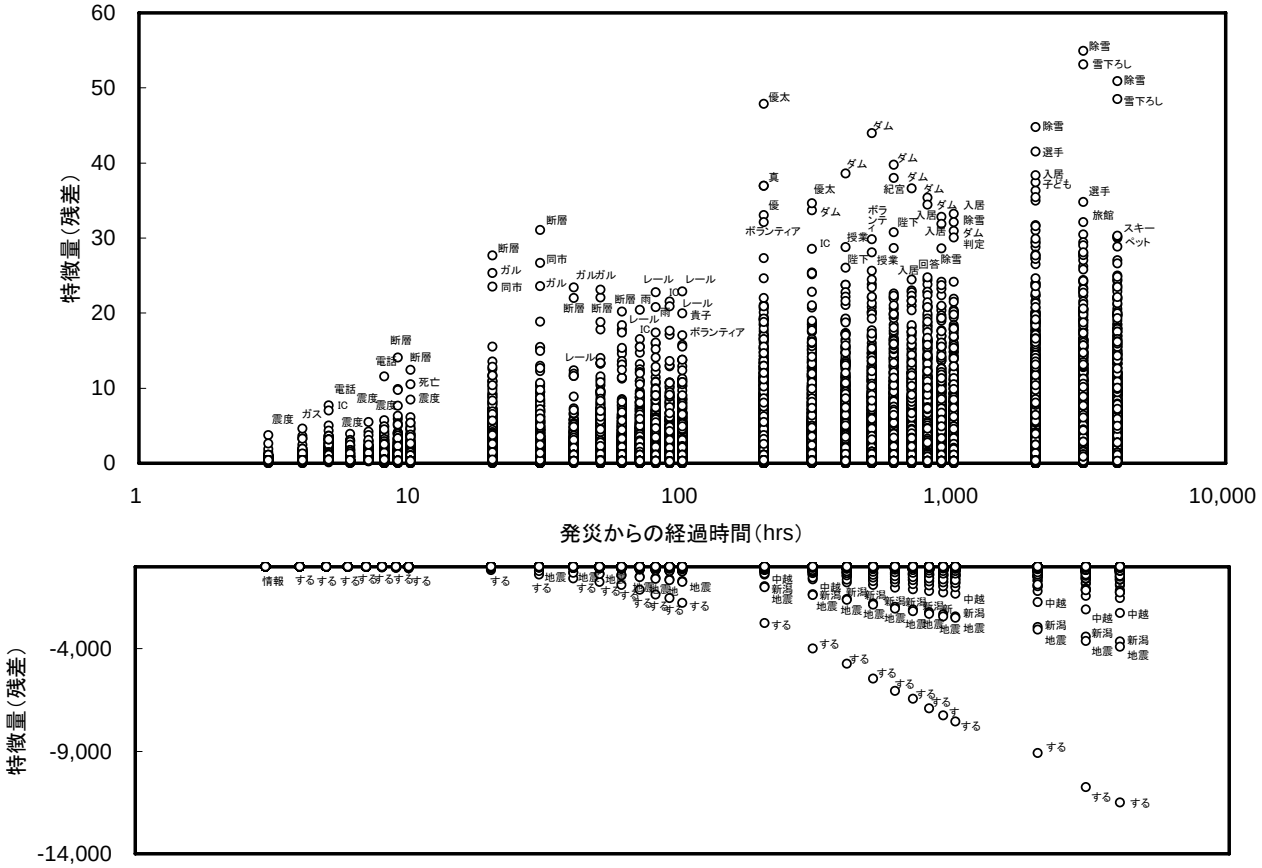

図 5　Σ 修正 TFIDF と ΣTF の関係にもとづく残差の評価結果

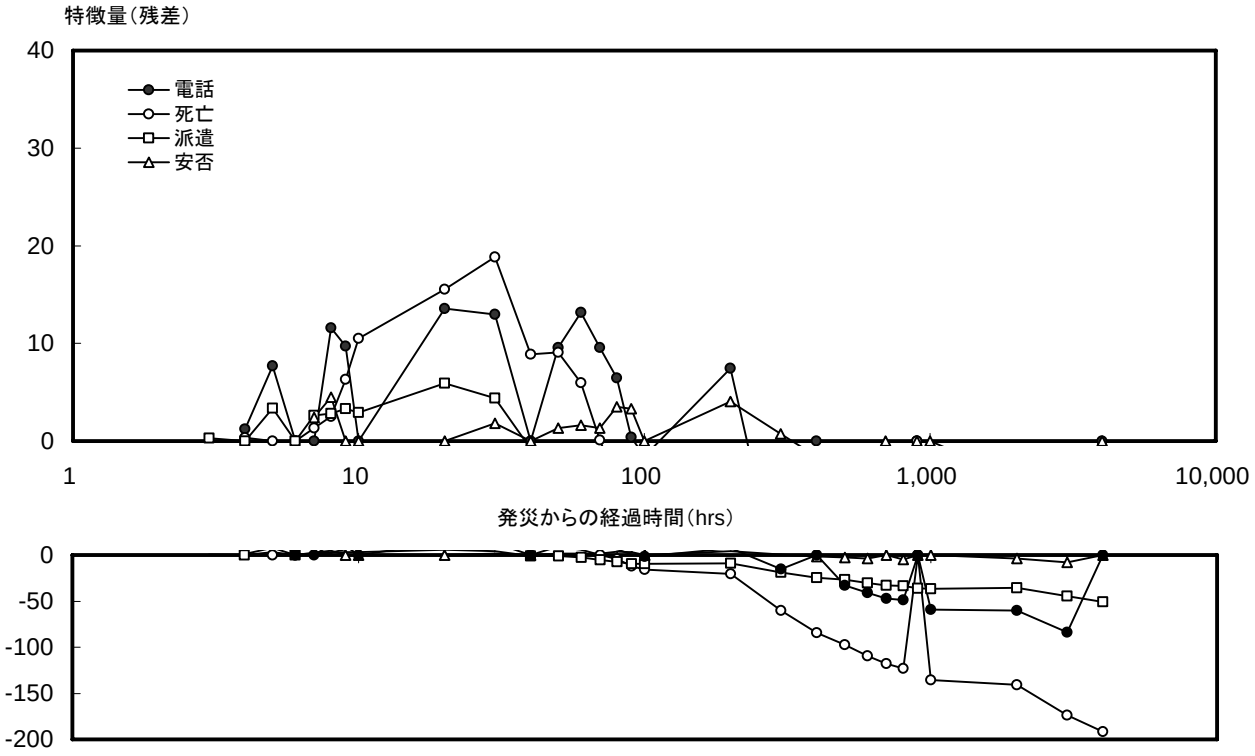

図 6　命を守る活動に関連するキーワードの特徴量の変化

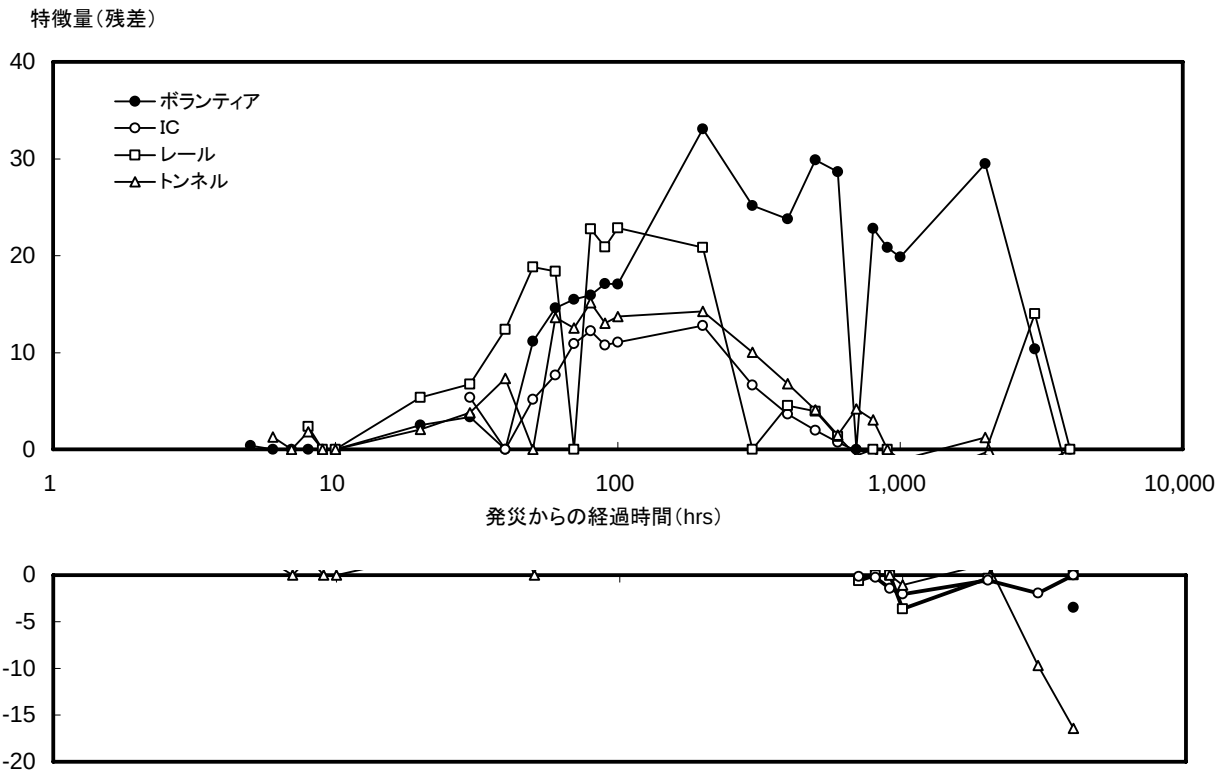

図 7　社会のフローを回復させる活動に関連するキーワードの特徴量の変化

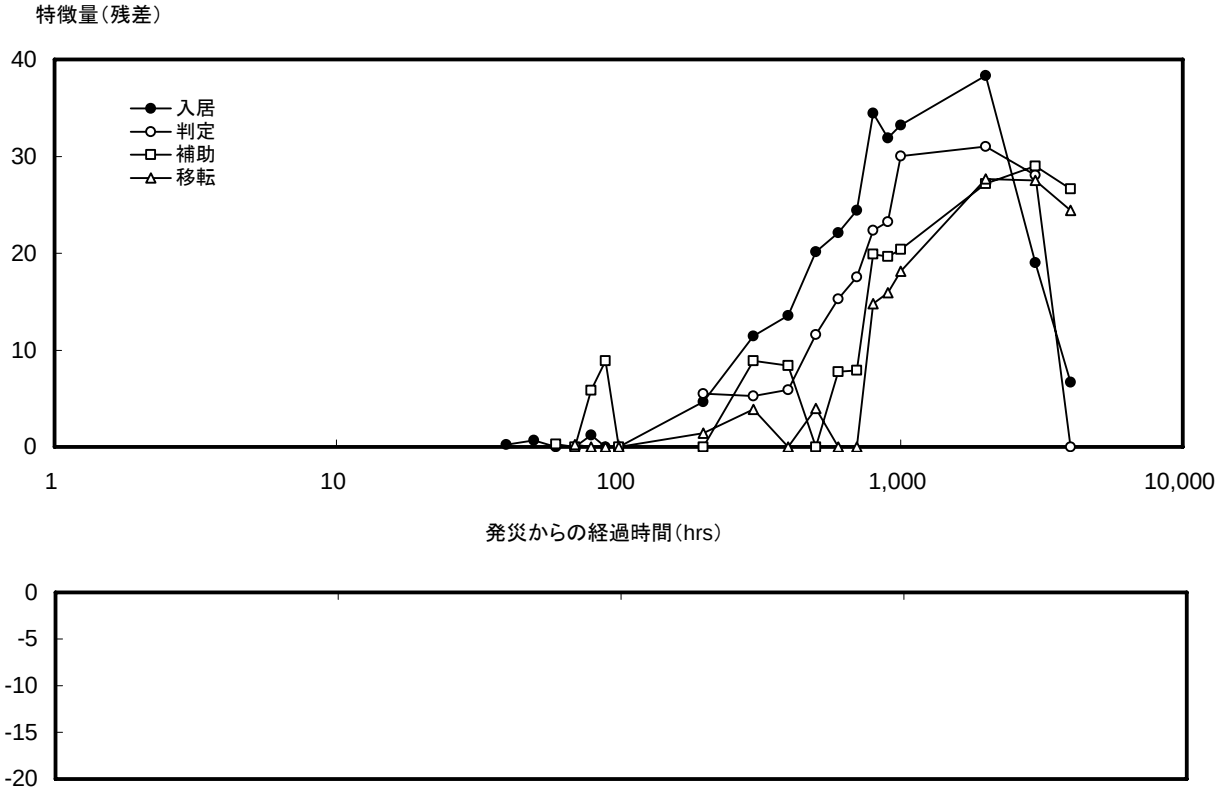

図 8　社会のストックを再建させる活動に関連するキーワードの特徴量の変化

表 1　特徴量が各時間断面で上位 10 位以上になったキーワードの一覧

キーワード		キーワード	
ボランティア	14	排水	4
IC	13	回答	4
断層	11	道路	3
震度	9	自宅	3
ダム	9	山	3
児童	9	義援金	3
レール	7	燕三条	3
電話	6	屋台	3
起きる	6	選手	3
同市	6	雪下ろし	3
トンネル	6	防災	2
雨	6	派遣	2
組合	6	安否	2
入居	6	発生	2
死亡	5	現在	2
羽田	5	県内	2
授業	5	震源	2
湖	5	小国	2
子ども	5	トイレ	2
判定	5	貴子	2
除雪	5	保険	2
補助	4	優	2
余震	4	陛下	2
土砂崩れ	4	大人	2
今回	4	紀宮	2
可能	4	補強	2
ガル	4	募金	2
加速度	4	業者	2
星野	4	旅館	2
村民	4	ペット	2
優太	4	移転	2

*数値は特徴量が上位10位以上になった回数

した．「入居（記事の例：山古志村の被災住民が 10 日午前，長岡市に建設された仮設住宅への入居を始めた(12/10 18:28 毎日新聞))」「判定(記事の例：建物の被害判定では 20 世帯が「不満だ」と回答(12/24 0:05 読売新聞))」などのすまいの再建に関するキーワードとなっている．これらのキーワードは，震災後 1,000 時間に特徴量が最も高くなる．また，社会にフローを回復させる活動とともに，社会のストックを再建するキーワードについては，それぞれ特徴量がピークとなる 100～1,000 時間，1,000 時間以降でキーワードが初出するわけではなく，それよりも早い時期に出現していた．

残差が正であったキーワードに対する以上のような考察から，1995 年に発生した阪神・淡路大震災の被災地でのエスノグラフィー調査や 2001 の米国 WTC テロ事件を取り上げたニュース記事に関する言語解析の結果にもとづく災害過程の理論によって想定されるキーワードが時間フェーズの層ごとに特徴的に検出されており，2004 年新潟県中越地震災害のウェブニュースを用いた解析結果においても，10 のべき乗の時間を節目として災害過程の質が変化するという災害過程のモデルとの整合が確認された．また，**図 6～8** に示したキーワード群は，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に対応するフェーズに特徴量のピーク時点をもつものの，ピーク時点の前後を中心として，解析対象の期間中に特徴量が少なからず観測されており，それぞれの災害対応の内容が時間の経過ともに変化していくのではなく，それぞれの活動のピークをもちながら，平行して展開していくという災害対応の時間的展開モデル [10] に符合している．

図 6～8 で示さなかったキーワードの中でも，**図 5** の上で高い特徴量を示しているものがある． 100～1,000 時間では，「ダム（記事の例：山古志村の芋川に大量の土砂が流れ込んでつくられた天然の「ダム湖(天然ダム)」が，1 日夜から 2 日にかけての降雨で満水に近い状態になった(11/2 12:53 毎日新聞))」が最も特徴的である．これは，前のフェーズで特徴的だった「雨」が被災地で発生し，天然ダムの決壊の危険性が高まったことにより，特徴量が高まったものと考えられる．被災地が豪雪地帯であったこと，当時は例年に比べて積雪量が多かったこと，屋根への積雪により地震で強度が低下した家屋が倒壊する危険性があったことからこの時期（1～3 月）「除雪」「雪下ろし」というキーワードも特徴的だった．これに伴い，除雪活動を支援する活動に関する「ボランティア」というキーワードの特徴量も再び高くなる．新潟県中越地震の場合には「ダム」「排水」「除雪」「雪下ろし」が検出されたように，本震以降に発生した降雨による土砂災害への影響や豪雪による建物倒壊の危険性という，地震動以外の自然ハザードによる二次災害の影響が特徴的に取り上げられていたことが明らかになった．

「同市」「今回」「可能」のように一部，キーワードとして適切でないと思われる言葉が検出されるものの，**図 5～8** や **表 1** にもとづく上述の考察のように，災害発生から復興までの各フェーズを代表するようなキーワードが検出されたことから，おおむね各言語資料（ニュース記事）の情報内容を表すキーワードの検出が可能になったことを確認できた．

また，**図 5** における残差が負であった語には，「する」「新潟」「地震」「中越」などが現れた．「する」のような日本語の語彙特性上，どのような文章に対しても使用頻度が高いと思われる語のほか，「新潟」「地震」「中越」など，ここでの解析に用いた災害の名称（新潟県中越地震）に含まれたキーワードが著しい残差の低さを示した．一般的に，危機の名称には，危機が発生した地域やハザードの名称が含まれることから，様々な危機に関する言語資料を収集して，本手法を適用した場合によって残差が著しく低い負の値で検出された地域名やハザード名のキーワードを「呼び出しタグ」とすることによって，言語資料体の中から異質なテキストデータの混入を検出することも可能である．

キーワードの特徴量を用いた**図 5～8** のようなかたちで可視化をおこなえば，本来大量のテキストで構成される言語資料をキーワードを単位として時系列的に情報の縮約を図ることができる．キーワードの時系列的な特徴の変化を XMDB のユーザーに提示することは，災害の過程の概況の大まかな理解を促し，データベースに蓄積さ

れている言語資料体からデータや情報，知識や教訓を得ようとするさいの検索キーワードの選定を補助する役割を担う．また，災害が発生している中で収集された言語資料に対して，開発したテキストマイニング手法をリアルタイムに適用すれば，大規模な量の言語情報が客観的・定量的に情報が集約され，実務者間など状況の認識の統一を図ることが可能となり，政策判断や意思決定を支援することが可能であると考えられる．

６．おわりに

この研究では，危機管理分野に関する言語資料体の適切なデータベース管理や知識・教訓の抽出をする容易な情報検索の実現を目的として，言語資料体からキーワードを自動的に検出する手法の構築を試みた．この研究の成果は以下のようにまとめられる：

1) 形態素解析，時間の概念を考慮した *TFIDF* 指標の算出，残差分析手法からなる言語資料体からキーワードを自動的・客観的に抽出する手法を開発した．
2) 本手法に 2004 年新潟県中越地震について取り上げたウェブニュースを適用した結果，各時間断面で固有性の高いキーワード（特徴語）と，ハザード名や被災地域名などの災害の発生から一貫して出現する共通性の高いキーワード（一般語）を判別して検出できることできた．
3) 2）で検出したキーワードと算出された特徴量を用いて言語資料体のもつ情報を縮約し，情報を縮約する時系列的なキーワードの特徴の変化を可視化するテキストマイニング手法を考案した．
4) 3）で構築したテキストマイニング手法を新潟県中越地震に関するウェブニュースに適用した結果，発災からの発信されたニュース記事の特徴の変化の概況をキーワードを単位として捉えることができた．
5) 新潟県中越地震に関するウェブニュースのテキストマイニング手法を適用した解析結果は，命を守る活動，社会のフローを回復させる活動，社会のストックを再建する活動が 10～100 時間，100～1,000 時間，1,000 時間以降を中心として，同時並行で展開するという災害過程のモデルと符合することが確認され，手法の一定の有効性が示された．

本研究は，新潟県中越地震について報道されたウェブニュースに対する手法の適用と分析にとどまっている．社会現象としての災害を記録する言語資料には，ウェブニュース以外にも体験談記録，災害対応の記録，被害報告，刊行資料などの言語資料などがある．これらはデジタル媒体で整備されていないこと，データフォーマットが資料や記録，情報源によって異なるなど，データの収集や蓄積の時点での課題を残している．ここに挙げた言語資料に加え，ウェブニュースを含むウェブ上の資料について，データの収集から蓄積から，キーワード付与や情報縮約に至るまでの過程の業務フローを標準化したい．

XMDB では，防災実務者や防災研究者が災害や危機のデータや研究成果から有用な情報，知識，教訓などの検索や抽出が容易に行える環境の構築を目指している．ユーザーの意向を反映できる情報検索は，ここで開発した手法によって検出されるキーワードとその特徴量や順位を利用した検索アルゴリズムを構築することによって実現できる可能性がある．今後は，算出された特徴量にもとづく適切な上位数を定めること，特徴量や順位の分析をおこない，ユーザビリティーの高い情報検索が行えるような検索アルゴリムの開発を課題とする．

これらの観点から，XMDB の機能性の拡張や蓄積データの充実を図り，危機管理に関する実務者や研究者にとって減災に寄与する情報．知識，教訓を抽出・共有できる環境の構築につなげていきたい．

謝辞

本研究は，文部科学省大都市大震災軽減化プロジェクトⅢ-3 第 5 課題「新公共経営（New Public Management）の枠組みにもとづく地震災害対応シミュレーターによる災害対応力の向上」（研究代表者：林春男 京都大学）および文部科学省科学技術振興調整費 先導的研究等の推進「日本社会に適した危機管理システム基盤構築」（研究代表者：林春男 京都大学）によるものである．

補 注

(1) コーパスとは，言語分析のための文字言語，あるいは音声言語資料の集合体として定義されるもので [11]，特に電子テキストで構築されたものを指す．一般には，電子的なオリジナルのテキスト群を収集したものを指すが，この研究では，上記の定義を広義にとらえ，オリジナルテキストに対して *TFIDF* や *TF* の情報をもつ形態素群を便宜的にコーパスと呼ぶことにする．

参考文献

1) 吉冨望，浦川豪，下田渉，川方裕則，林春男：防災情報共有のためのクロスメディアデータベースの構築，地域安全学会論文集，No. 6, pp. 315-322, 2004.
2) 亀田弘行：平成 7 年兵庫県南部地震をふまえた大都市災害に対する総合防災対策の研究，文部科学省緊急プロジェクト，37pp. , 1995.
3) 松村劭：オペレーショナル・インテリジェンス 意思決定のための作戦情報理論，日本経済新聞社，220pp. , 2006..
4) Yahoo Japan News：http://headlines.yahoo.co.jp/hl
5) 北研二，津田和彦，獅子堀正幹：情報検索アルゴリズム，共立出版，212pp. , 2002.
6) 松本裕治：形態素解析システム「茶筌」，情報処理，Vol. 41, No. 11, pp. 1208-1214, 2000.
7) 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.
8) 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.
9) 立木茂雄：被災者支援原則の構築，文部科学省科学技術振興調整費「日本社会に適した危機管理システム基盤構築」研究成果発表ワークショップ アブストラクト集，pp. 37-38, 2006.
10) 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.
11) 伊藤雅光：計量言語学入門，大修館書店，285pp. , 2002.

(原稿受付　2006.5.26)
(登載決定　2006.9.16)