(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4913154号
(P4913154)
(45)発行日　平成24年4月11日(2012.4.11)　(24)登録日　平成24年1月27日(2012.1.27)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 210A
G06F 17/30 170A

請求項の数 10　(全 35 頁)

(21)出願番号　特願2008-545465 (P2008-545465)
(86)(22)出願日　平成19年11月22日 (2007.11.22)
(86)国際出願番号　PCT/JP2007/073257
(87)国際公開番号　WO2008/062910
(87)国際公開日　平成20年5月29日 (2008.5.29)
審査請求日　平成21年7月17日 (2009.7.17)
(31)優先権主張番号　特願2006-315238 (P2006-315238)
(32)優先日　平成18年11月22日 (2006.11.22)
(33)優先権主張国　日本国(JP)

(73)特許権者　506391440
林　春男
京都府宇治市折居台１－４－２９
(74)代理人　100090181
弁理士　山田　義人
(72)発明者　林　春男
京都府宇治市折居台１－４－２９

審査官　野崎　大進

最終頁に続く

(54)【発明の名称】文書解析装置および方法

(57)【特許請求の範囲】
【請求項１】
時系列的に増量する言語資料を解析する文書解析装置であって、
時系列順序を有し、かつ前記時系列順序が後のものが先のものに比べて多い数の単位ドキュメントのテキストデータを含むテキストコーパスを作成するコーパステキスト作成手段、
前記コーパステキストに含まれるテキストデータを構成する形態素に品詞情報を付加する形態素解析手段、
前記品詞情報に基づいて前記テキストデータから不要な形態素を取り除く不要形態素除去手段、
前記不要形態素除去手段によって除去されなかった形態素について、形態素毎に、時間増加型ＴＦＩＤＦを計算して時間増加型ＴＦＩＤＦの実測値を得る計算手段、および
前記計算手段で計算した前記実測値の累計値と前のコーパスにおいて推定した前記時間増加型ＴＦＩＤＦの累計値の推定値との間で残差分析をして形態素毎の残差値を求める残差分析手段を備える、文書解析装置。
【請求項２】
各コーパスにおいて，任意時点のコーパスから求められる形態素毎の時間増加型ＴＦＩＤＦの累計値とＴＦの累計値とで回帰曲線を作成する回帰曲線作成手段をさらに備え、
前記残差分析手段は、前記回帰曲線作成手段が前の時点のコーパスで作成した回帰曲線と、現在のコーパスにおいて前記計算手段が計算した各形態素の前記時間増加型ＴＦＩＤ

Ｆの前記実測値との間で残差分析を行なう、請求項１記載の文書解析装置。
【請求項３】
前記残差分析手段による残差分析の結果、正の残差値が得られた形態素を当該コーパスにおける特異語として選定する特異語選定手段をさらに備える、請求項１または２記載の文書解析装置。
【請求項４】
前記特異語選定手段は、フィルタリング処理を実行するフィルタリング手段を含む、請求項３記載の文書解析装置。
【請求項５】
前記特異語選択手段によって選択した特異語を可視的に出力する特異語出力手段をさらに備える、請求項３または４記載の文書解析装置。
【請求項６】
前記残差分析手段による残差分析の結果、負の残差値が得られた形態素を当該コーパスの共通語として選定する共通語選定手段をさらに備える、請求項１ないし５のいずれかに記載の文書解析装置。
【請求項７】
前記共通語選択手段によって選択した共通語を可視的に出力する共通語出力手段をさらに備える、請求項６記載の文書解析装置。
【請求項８】
前記特異語出力手段によって出力された特異語の少なくとも１つについて、当該特異語が含まれる単位ドキュメントを可視的に出力するドキュメント出力手段をさらに備える、請求項５記載の文書解析装置。
【請求項９】
時系列的に増量する言語資料を解析する文書解析装置のコンピュータによって実行される文書解析プログラムであって、前記コンピュータを
時系列順序を有し、かつ前記時系列順序が後のものが先のものに比べて多い数の単位ドキュメントのテキストデータを含むコーパステキストを作成するコーパステキスト作成手段、
前記コーパステキストに含まれるテキストデータを構成する形態素に品詞情報を付加する形態素解析手段、
前記品詞情報に基づいて前記テキストデータから不要な形態素を取り除く不要形態素除去手段、
前記不要形態素除去手段によって除去されなかった形態素について、形態素毎に、時間増加型ＴＦＩＤＦを計算して時間増加型ＴＦＩＤＦの実測値を得る計算手段、および
前記計算手段で計算した前記実測値と前のコーパスにおいて推定した前記時間増加型ＴＦＩＤＦの累計値の推定値との間で残差分析をして形態素毎の残差値を求める残差分析手段
として機能させる、文書解析プログラム。
【請求項１０】
時系列的に増量する言語資料を解析する文書解析装置のコンピュータが実行する文書解析方法であって、
時系列順序を有し、かつ前記時系列順序が後のものが先のものに比べて多い数の単位ドキュメントのテキストデータを含むコーパステキストを作成するコーパステキスト作成ステップ、
前記コーパステキストに含まれるテキストデータを構成する形態素に品詞情報を付加する形態素解析ステップ、
前記品詞情報に基づいて前記テキストデータから不要な形態素を取り除く不要形態素除去ステップ、
前記不要形態素除去ステップによって除去されなかった形態素について、形態素毎に、時間増加型ＴＦＩＤＦを計算して時間増加型ＴＦＩＤＦの実測値の累計値を得る計算ステ

ップ、および
前記計算ステップで計算した前記実測値の累計値と前のコーパスにおいて推定した前記時間増加型ＴＦＩＤＦの累計値の推定値との間で残差分析をして形態素毎の残差値を求める残差分析ステップを含む、文書解析方法。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は文書解析装置および方法に関し、特にたとえばニュース，ウェブニュース，ブログ，新聞，雑誌，インタビュー記録，供述調書，アンケート，小説などのように、時系列的に増量する言語資料から時系列順序に応じた特異語（キーワード）を抽出または検出できる、新規な文書解析装置および方法に関する。
【背景技術】
【０００２】
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。
【０００３】
このような防災に関する情報は、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持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との間のコミュニケーションも困難なものになっている。
【０００４】
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、ユーザが親しみやすいインタフェースを使って、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。
【０００５】
発明者等は、防災研究者や防災実務者の間で情報を共有しまたは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」という。）の開発を試みてきている（非特許文献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）。
【０００６】
このＸＭＤＢに蓄積すべきデータや情報は、強震計による揺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。
【０００７】
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（非特許文献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）。
【０００８】
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。

【０００９】

自然災害科学の領域では、強震計による揺れの観測結果、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行ない、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、シミュレーションの入力外力として用いられたりして、構造物の耐力向上に資する研究がなされている。

【００１０】

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行ない教訓や知識を抽出し体系化し、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。

【００１１】

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際には、以下のような問題が発生する。

【００１２】

まず１点目として、データベースへの蓄積の際、各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，テーマに基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。

【００１３】

こ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（非特許文献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）。

【００１４】

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。

【００１５】

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。

【００１６】

一方、文書データからキーワードを抽出する方法が特許文献1（特開2004‐5711号公報[G06F 17/30]）などで提案されている。

【００１７】

特許文献１のキーワード抽出装置および方法は、固定的に定まった量の文書を対象にしているため、たとえばニュースなどのように、時系列的に順序を有し、あるいは時系列的に情報量が増加する性質を持つテキストデータ群に有効に対処できない。

【非特許文献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

【特許文献１】特開2004‐5711号公報 [G06F 17/30]
【発明の概要】
【発明が解決しようとする課題】
【００１８】
それゆえに、この発明の主たる目的は、新規な、文書解析装置および方法を提供することである。
【００１９】
この発明の他の目的は、時系列的に増量する言語資料から適切な特異語（キーワード）と共通語を検出できる、文書解析装置および方法を提供することである。
【課題を解決するための手段】
【００２０】
この発明は、上記の課題を解決するために、以下の構成を採用した。なお、括弧内の参照符号および補足説明等は、この発明の理解を助けるために後述する実施形態との対応関係を示したものであって、この発明を何ら限定するものではない。
【００２１】
第１の発明は、時系列的に増量する言語資料を解析する文書解析装置であって、時系列順序を有し、かつ時系列順序が後のものが先のものに比べて多い数の単位ドキュメントのテキストデータを含むテキストコーパスを作成するテキストコーパス作成手段、コーパステキストに含まれるテキストデータを構成する形態素に品詞情報を付加する形態素解析手段、品詞情報に基づいてテキストデータから不要な形態素を取り除く不要形態素除去手段、不要形態素除去手段によって除去されなかった形態素について、形態素毎に、時間増加型ＴＦＩＤＦを計算して時間増加型ＴＦＩＤＦの実測値を得る計算手段、および計算手段で計算した実測値と前のコーパスにおいて推定した時間増加型ＴＦＩＤＦの累計値の推定値との間で残差分析をして形態素毎の残差値を求める残差分析手段を備える、文書解析装置である。
【００２２】
第１の発明では、文書解析装置は、典型的には、コンピュータで構成される。コーパステキスト作成手段（Ｓ３：実施例で対応する部分を例示的に示す参照符号。以下同様。）は、たとえば予め設定した時間が経過すると、時系列順序が先のコーパスに比べて、含まれる単位ドキュメントの数が多い現在時間のコーパスを作成する。時間経過とともに逐次増量するたとえばウェブニュースのような場合には、設定時間（設定時間は任意である。）の経過に伴ってそのウェブニュースのテキストデータを用いてコーパステキストを作成するが、言語資料には逐次増量する文書だけでなく、単に時系列順序だけを有する文書もある。後者の場合には、コーパス作成手段は時間経過に応じてコーパステキストを順次作成するのではなく、時系列順序に先後のある複数のコーパステキストを一度に準備または作成するようにしてもよい。
【００２３】
形態素解析手段（Ｓ５）は、たとえば日本語のように形態素が分割されていない言語体系のテキストデータである場合、たとえば茶筅（http://chasen.naist.jp/hiki/ChaSen/）のような形態素解析ツールを用いて、そのコーパスに含まれる単位ドキュメントのテキストデータを形態素に分解して、各形態素に品詞情報を付加する。しかしながら、テキスト内の形態素が既に分割している、たとえば英語のような言語体系の場合には、形態素を分割する作業は必要ではなく、この形態素解析手段では、たとえばタギング処理によって、テキストを構成する各形態素に品位情報を付加する。
【００２４】
不要形態素除去手段（Ｓ７）は、各形態素に付加された上述の品詞情報に基づいて、不要形態素として設定しておいた品詞の種類の形態素を取り除く。つまり、形態素解析の際に、各形態素に付与される品詞情報に基づいて、当該形態素を特異語および／または共通語の候補として採用するか否かを選定する。ただし、不要とする形態素の品詞の種類は、任意に設定できる。

【００２５】
計算手段（Ｓ１１）は、そのコーパスに残った形態素の各々について、ＴＦ（Term Frequency）つまり単位ドキュメント中にそのキーワード候補が出現する頻度(延べ数)を計算し、さらに時間のパラメータを考慮したＩＤＦ（Inversed Document Frequency）つまり他には出現していないという独自性値を計算することによって、当該コーパスにおける当該形態素の時間増加型ＴＦＩＤＦ（Term Frequency Inversed Document Frequency）を「ＴＦ」×「ＩＤＦ」として計算する。
【００２６】
残差分析手段（Ｓ１７）は、たとえば、時間的順序が前のコーパスにおいて推定しておいた該当の形態素の時間増加型ＴＦＩＤＦの累計値の推定値と、上記計算手段が計算した時間増加型ＴＦＩＤＦの累計値の実測値との間で残差分析を行ない、その形態素の残差値（正，負）を求める。
【００２７】
第１の発明によれば、言語資料体が時系列的に増量するものであっても、コーパス作成手段が、時系列順序が後のものが先のものに比べて多い数の単位ドキュメントを含むテキストコーパスを作成し、それらコーパスに基づいて時間増加型ＴＦＩＤＦの累計値を目的変数とし、ＴＦの累計値を説明変数とする回帰曲線を作成していているため、現在のコーパスの時間増加型ＴＦＩＤＦの累計値を、その前のコーパスで作成された回帰曲線上に当該指標が分布するものと仮定して、現在のコーパスのＴＦの累計値を入力値とする現在のコーパスの時間増加型ＴＦＩＤＦの累計値の推定値を得るという処理の流れによって、その言語資料体を確実に解析することができる。
【００２８】
第２の発明は、第１の発明に従属し、各コーパスにおいてそのコーパスまでの時間増加型ＴＦＩＤＦの累計値とＴＦの累計値とで回帰曲線を作成する回帰曲線作成手段をさらに備え、残差分析手段は、回帰曲線作成手段が前の時点のコーパスで作成した回帰曲線と、現在の時点のコーパスにおいて計算手段が計算した各形態素の時間増加型ＴＦＩＤＦの累計値の実測値との間で残差分析を行なう、文書解析装置である。
【００２９】
第２の発明では、回帰曲線作成手段は、説明変数であるＴＦの累計値（ΣＴＦ）をＸとし、従属変数である時間増加型ＴＦＩＤＦの累計値（Σ時間増加型ＴＦＩＤＦ）をＹとして、定数を計算して回帰曲線を作成する。ただし、このような回帰曲線の計算は、時系列順序が前のコーパスで予め計算しておくものである。第２の発明によれば、時系列順序が前のコーパスにおいて時系列順序が後のコーパスにおける時間増加型ＴＦＩＤＦの累計値の推定または予測のための回帰曲線を準備しておくので、当該後のコーパスにおける残差分析が迅速に行なえる。
【００３０】
第３の発明は、第１または第２の発明に従属し、残差分析手段による残差分析の結果、正の残差値が得られた形態素を当該コーパスにおける特異語として選定する特異語選定手段をさらに備える、文書解析装置である。
【００３１】
第３の発明では、特異語選定手段（Ｓ２１，Ｓ２１Ａ，Ｓ２１Ｂ）が、正の残差値（の大きなもの）を有する形態素を特異語として選定する。第３の発明によれば、残差値だけをパラメータとして選定するので、客観的な特異語が選定できる。その特異語は、当該コーパスの特徴を表すキーワードとして機能する。
【００３２】
第４の発明は、第３の発明に従属し、特異語選定手段は、フィルタリング処理を実行するフィルタリング手段を含む、文書解析装置である。
【００３３】
第４の発明では、コンピュータ（１４）は、ユーザが選択的にフィルタリングをオプションとして設定した場合、たとえば、（１）Δｔにおいて出現文書数が１件である語（形

態素）を除外するというフィルタリング１および／またはたとえば（２）出現文書数と語（形態素）の出現頻度との関係から、出現頻度が著しく高い形態素を除外するというフィルタリング２を実行する。それによって、極端に高い特異値を示す形態素を除外することができる。

【００３４】

第５の発明は、第３または第４の発明に従属し、特異語選択手段によって選択した特異語を可視的に出力する特異語出力手段をさらに備える、文書解析装置である。

【００３５】

第５の発明では、コンピュータ（１４）は、たとえば図１５-図２１および図２７－図２９に示すように、特異語選定手段が設定した特異語をたとえばグラフ形式で可視化表示（出力）する。

【００３６】

第６の発明は、第１ないし第５の発明いずれかに従属し、残差分析手段による残差分析の結果、負の残差値が得られた形態素を当該コーパスの共通語として選定する共通語選定手段をさらに備える、文書解析装置である。

【００３７】

第６の発明では、共通語選定手段（Ｓ２１）が、負の残差値（の大きなもの）を有する形態素を共通語として選定する。第６の発明によれば、残差値だけをパラメータとして選定するので、客観的な共通語が選定できる。その共通語は、当該コーパスだけでなく他のコーパスをグループ化するためのインデックスなどとして機能する。

【００３８】

第７の発明は、第６の発明に従属し、共通語選択手段によって選択した共通語を可視的に出力する共通語出力手段をさらに備える、文書解析装置である。

【００３９】

第７の発明では、コンピュータ（１４）は、たとえば図１５-図２１に示すように、共通語選定手段が設定した共通語をたとえばグラフ形式で可視化表示（出力）する。

【００４０】

第８の発明は、第５の発明に従属し、特異語出力手段によって出力された特異語の少なくとも１つについて、当該特異語が含まれる単位ドキュメントを可視的に出力するドキュメント出力手段をさらに備える、文書解析装置である。

【００４１】

第８の発明では、たとえば各時点で作成された形態素（ｔｉ）の特異値（ＤＶｔｉ）リストに基づいて、今回のコーパスに含まれる単位ドキュメントごとに、その単位ドキュメントに含まれる特異語（特異値が高い上位１０の語）について、特異値の総和を求める。特異値の総和（ＲＶ）の高い、少なくとも１つの単位ドキュメント（文書）をたとえば「注目記事」として選定し、その選定した単位ドキュメントをたとえばテキストデータテーブル（２０）から読み出して、少なくとも見出しを、その特異語とともに表示する。第８の発明によれば、特異値の総和が大きい語（形態素）を含む単位ドキュメント（記事）の少なくとも見出しが、必要に応じて本文も含めて、表示される。そのため、解析によって失われた形態素の文脈の情報を補完でき、高い特異性を示した形態素の理解や解釈を容易にする。

【００４２】

第９の発明は、時系列的に増量する言語資料を解析する文書解析装置のコンピュータによって実行される文書解析プログラムであって、コンピュータを、時系列順序を有し、かつ時系列順序が後のものが先のものに比べて多い数の単位ドキュメントのテキストデータを含むコーパステキストを作成するコーパステキスト作成手段、コーパステキストに含まれるテキストデータを構成する形態素に品詞情報を付加する形態素解析手段、品詞情報に基づいてテキストデータから不要な形態素を取り除く不要形態素除去手段、不要形態素除去手段によって除去されなかった形態素について、形態素毎に、時間増加型ＴＦＩＤＦを計算して時間増加型ＴＦＩＤＦの実測値を得る計算手段、および計算手段で計算した実測

値と前のコーパスにおいて推定した時間増加型ＴＦＩＤＦの累計値の推定値との間で残差分析をして形態素毎の残差値を求める残差分析手段として機能させる、文書解析プログラムである。

【００４３】

第１０の発明は、時系列的に増量する言語資料を解析する文書解析装置のコンピュータが実行する文書解析方法であって、時系列順序を有し、かつ時系列順序が後のものが先のものに比べて多い数の単位ドキュメントのテキストデータを含むコーパステキストを作成するコーパステキスト作成ステップ、コーパステキストに含まれるテキストデータを構成する形態素に品詞情報を付加する形態素解析ステップ、品詞情報に基づいてテキストデータから不要な形態素を取り除く不要形態素除去ステップ、不要形態素除去ステップによって除去されなかった形態素について、形態素毎に、時間増加型ＴＦＩＤＦを計算して時間増加型ＴＦＩＤＦの実測値の累計値を得る計算ステップ、および計算ステップで計算した実測値の累計値と前のコーパスにおいて推定した時間増加型ＴＦＩＤＦの累計値の推定値との間で残差分析をして形態素毎の残差値を求める残差分析ステップを含む、文書解析方法である。

【００４４】

第９の発明および第１０の発明は、基本的に第１の発明と同様である。

【発明の効果】

【００４５】

この発明によれば、言語資料の増量に応じて、時系列順序が先後のコーパスにおいて単位ドキュメントの数を増加させたコーパスを作成するようにしているので、言語資料が時系列的に増量するものであっても、確実に分析または解析して、たとえば特異語や共通語を抽出することができる。

【００４６】

この発明の上述の目的，その他の目的，特徴，および利点は、図面を参照して行う以下の実施例の詳細な説明から一層明らかとなろう。

【図面の簡単な説明】

【００４７】

【図１】図１はこの発明の一実施例であるキーワード検出システムを示すブロック図である。

【図２】図２はこの実施例で用いられるテキストデータテーブルの一例を示す図解図である。

【図３】図３は図１実施例のコンピュータの動作を示すフロー図である。

【図４】図４はこの実施例で作成する時間とともに増加するコーパスの一例を示す図解図である。

【図５】図５は各記事および形態素の出現頻度の解析結果の一例を示す表である。

【図６】図６は各記事および形態素に対する単位ドキュメント数Ｎを示す表であり、図６（Ａ）が言語資料体が一定量である一般的な場合（時間の経過とともに増加しない場合）を示し、図６（Ｂ）が時系列的に増量する言語資料体を解析する実施例の場合を示す。（Ａ）は，他の図（図５～８）との表記を統一させるために，表示例の形態素（t1，t2，t3・・・）毎に単位ドキュメント数Ｎを示してある。

【図７】図７は各記事および形態素に対するＤＦを示す表であり、図７（Ａ）が言語資料体が一定量である一般的な場合（時間の経過とともに増加しない場合）を示し、図７（Ｂ）が時系列的に増量する言語資料体を解析する実施例の場合を示す。

【図８】図８は各記事および形態素に対するＴＦＩＤＦ（Ａ）および時間増加型ＴＦＩＤＦ（Ｂ）を示す表であり、図８（Ａ）が言語資料体が一定量である一般的な場合（時間の経過とともに増加しない場合）を示し、図８（Ｂ）が時系列的に増量する言語資料体を解析する実施例の場合を示す。

【図９】図９は回帰曲線の一例を示す図解図である。

【図１０】図１０は回帰曲線とそれに対する残差（正負）を示すグラフであり、横軸にＴ

Ｆの総和を、縦軸に時間増加型ＴＦＩＤＦの総和をとる。
【図１１】図１１は図１実施例のコンピュータで表示される１つの表示例を示す図解図である。
【図１２】図１２は図１実施例のコンピュータで表示される別の表示例を示す図解図である。
【図１３】図１３はコーパス毎の図９と同様の回帰曲線を示すグラフであり、図１３(Ａ)が発災から１０時間後のコーパスにおける回帰曲線を示し、図１３(Ｂ)が発災から１００時間後のコーパスにおける回帰曲線を示し、図１３(Ｃ)が発災から１０００時間後のコーパスにおける回帰曲線を示し、図１３(Ｄ)が発災から４５００時間後のコーパスにおける回帰曲線を示す。
【図１４】図１４はコーパスと回帰曲線との関係を示す図解図である。
【図１５】図１５は図１実施例を用いて実際のウェブニュースから求めた発災から１０時間内の特徴量（上が正、下が負）を示す図解図である。
【図１６】図１６は図１５と同様にして求めた発災から１０－１００時間内の特徴量を示す図解図である。
【図１７】図１７は図１５と同様にして求めた発災から１００－５００時間内の特徴量を示す図解図である。
【図１８】図１８は図１５と同様にして求めた発災から５００－１０００時間内の特徴量を示す図解図である。
【図１９】図１９は図１５と同様にして求めた発災から１０００－２０００時間内の特徴量を示す図解図である。
【図２０】図２０は図１５と同様にして求めた発災から２０００－３０００時間内の特徴量を示す図解図である。
【図２１】図２１は図１５と同様にして求めた発災から３０００－４５００時間内の特徴量を示す図解図である。
【図２２】図２２は図１実施例を用いて実際のウェブニュースから抽出したキーワードの変遷を示す図解図である。
【図２３】図２３はこの発明の他の実施例における図１のコンピュータの動作を示すフロー図である。
【図２４】図２４はこの他の実施例でメモリに記憶する、各語の出現頻度ＴＦと出現文書数ＤＦを示す図解図である。
【図２５】図２５はこの他の実施例における回帰直線と９５％信頼限界の一例を示すグラフである。
【図２６】図２６はこの他の実施例における回帰直線と９５％信頼限界の他の例を示すグラフである。
【図２７】図２７はフィルタリングオプションを選択しなかった場合の特異語のグラフ表示を示す図解図である。
【図２８】図２８はフィルタリング１をオプションとして選択した場合の特異語のグラフ表示を示す図解図である。
【図２９】図２９はフィルタリング２をオプションとして選択した場合の特異語のグラフ表示を示す図解図である。
【発明を実施するための最良の形態】
【００４８】
図１に示すこの発明の一実施例の文書解析装置１０は、たとえばインターネットのような通信網（ネットワーク）１２に有線または無線で結合されるコンピュータ１４を含む。コンピュータ１４には、基本的に、キーボードやマウスのような操作手段１５Ａおよび液晶表示器のようなモニタ１５Ｂが設けられていて、このコンピュータ１４には、さらに、テキストデータベース１６および分析データベース１８が付設される。コンピュータ１４は当然、内部メモリを有し、その内部メモリ（図示せず）はワーキングメモリなどとして利用され、計算して得られた結果データや、解析結果データ、さらにはその解析途中の各

種データなどを一時的に記憶する。

【００４９】

テキストデータベース１６には、たとえば、このコンピュータ１４がネットワーク１２を通して取得した時間順次のウェブニュースのテキストデータが逐次記憶され、コンピュータ１４はこのウェブニュースのテキストデータを順次分析または解析することによって、時系列的に変遷する特異語（キーワード）を抽出する。

【００５０】

テキストデータベース１６に蓄積されるテキストデータテーブル２０の一例が図２に示される。テキストデータテーブル２０は、具体的には、テキストデータで構成される言語資料から、任意の一定の大きさをもつ「単位ドキュメント」のテキストデータを１つのレコードに持つテーブルである。

【００５１】

単位ドキュメントの例としては、ウェブニュースの場合であれば、所定期間内の記事、１日の記事、１つの記事、１つの段落、１つの文などがある。新聞を例にとれば、１紙、１つの記事、１つの段落、１つの文などがある。文学作品（小説）などの場合には、１つの作品、１つの章、１つの段落、１つの文などがある。その他、ウェブ上のブログを解析対象とした場合には，１つの日記を単位ドキュメントとしたり、コールセンターへの１つの問い合わせや苦情などを単位ドキュメントにしたりするなど、言語資料に対して任意の単位を「単位ドキュメント」として定めて、データベース２０を作成する。

【００５２】

図２に示すように、1つのレコードに対しては、数度やアルファベットなどで形成される識別子（ＩＤ番号）２２およびテキストデータ２４のほか、時間情報（時刻スタンプ）２６をメタデータとして付与する。時間情報２６には、ウェブニュース記事であれば発信日時、コールセンターへの問い合わせであれば問い合わせ時間などが該当する。この実施例の文書解析装置１０は、ニュースやブログなど時間とともに文字数が増加していく言語情報を対象としている。しかしながら、文学作品等のように常には更新されないような言語資料であっても、言語資料は線状性を有しているため、言語資料を読む人は、時間の経過ともに言語情報を理解することになる。したがって、小説や文学作品のように一見静的で時間情報を持たない言語資料については、図２に示す時間情報２６のフィールドに、時間情報の代わりに順序情報（１章、２章…、１段落目、２段落目…、１文目、２文目…など）をメタデータとして付与すればよい。その他、必要に応じて任意のフィールド、たとえばタイトル２６を設けて、データベーステーブル２０を作成する。

【００５３】

もし、このテキストデータテーブル２０をコンピュータ１４が作成するときには、たとえばコンピュータ１４の中にインストールされている、ＤＢＭＳ（Data Base Management System：データベース管理システム）のようなアプリケーションを用いて、たとえばネットワーク１２を通して取得したウェブニュースなどからテキストデータテーブルを作成することができる。

【００５４】

なお、図２に示す１つの識別記号（ＩＤ）２２で区別されるかつ時系列情報２６が付された１つの単位ドキュメントのテキストデータ２４（図２）を含むものを、１レコードと呼ぶ。そして、言語資料体（コーパス）とは、このようなレコードの集合を意味する。

【００５５】

後述の実施例では、キーワード（特異語）を検出すべき時系列的に増量する言語資料体として、ウェブニュースを試用しているが、この種の言語資料としては、他に、新聞,雑誌,ブログ，インタビュー記録，供述調書，アンケート，小説など任意の時間要素を含むデータが想定できる。

【００５６】

分析データベース１８には、後述の形態素分析のための品詞辞書など、この実施例においてキーワード検出に必要な全ての辞書や文法ルールなどを予め記憶しているとともに、

分析結果も蓄積する。ただし、この分析データベース１８は、上述のテキストデータベース１６も同様であるが、コンピュータ１４の内部メモリで構成されていてもよい。

【００５７】

コンピュータ１４は、図３に示すキーワード抽出プログラムに従ってキーワードを抽出ないし検出する。

【００５８】

図３を参照して、最初のステップＳ１で、コンピュータ１４は、設定時間が経過したかどうか判断する。「設定時間」とは、時系列的に増量する言語資料から、時系列順序を有する各コーパスを画定するための、区切りの時間（Δｔ）である。この「設定時間」はユーザが自由に設定できる。たとえば、状況変化が短時間で生じるような言語資料を分析する際には、短い設定時間（Δｔ）を設定すればよく、逆の言語資料の場合には、設定時間Δｔを長くすればよい。Δｔの例としては、１時間、１０時間、１００時間、１日、１週間、１ヶ月など挙げられる。また、このΔｔを時間の経過とともに変更することも考えられる。一例として、災害発生から２４時間経過するまではたとえばΔｔを「１時間」に設定し、それ以降災害から３日目まではたとえばΔｔを「１０時間」に設定し、さらに１ヶ月以上経過したときにはたとえばΔｔを「１日」として設定する。

【００５９】

そして、ユーザによって任意の設定時間が設定されると、その設定時間はコンピュータ１４の適宜のメモリ領域（レジスタ）に記憶されるので、コンピュータ１４は、内部の時計データをレジスタに設定された設定時間と比較することによって、ステップＳ１で設定時間が経過したかどうか、判断することができる。

【００６０】

ステップＳ１で“ＹＥＳ”が判断されると、続いてコンピュータ１４はステップＳ３においてコーパス作成処理を実行し、設定時間（Δｔ）の間に増量した単位ドキュメントのテキストデータを、たとえば図２に示すテキストデータテーブル２０から読み込み、今回のテキストコーパスＣｔを作成する。

【００６１】

図４に示すコーパスＣｔは現在時間のコーパスを示すが、このコーパスＣｔは、それぞれより時系列順序が先のコーパスＣｔ-Δｔより、設定時間Δｔ後に形成したコーパスである。つまり、コーパスＣｔは、直前のコーパスＣｔ-Δｔと増量分のコーパスＣΔｔとを合計したものである。

【００６２】

なお、「コーパス（corpus）」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、特に電子テキストで構築されたものを指し、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、この実施例では、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時間増加型ＴＦＩＤＦやＴＦ（いずれも後述）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。したがって、ここでいうテキストコーパスは、少なくとも１つのレコードつまり少なくとも１つの単位ドキュメントのテキストデータを含む言語資料体を意味するものと理解されたい。

【００６３】

続いて、ステップＳ５において、そのコーパスに含まれるテキストデータ２４（図２）を形態素に分割し、品詞情報を付加する。ここで、形態素解析とは、自然言語で書かれた文を形態素(Morpheme、おおまかにいえば、言語で意味を持つ最小単位)の列に分割し、品詞を見分ける言語処理のことである。参照する情報源として、対象言語の文法の知識（ここでは文法のルールの集まり）と辞書（品詞等の情報付きの単語リスト）を用いるが、これらの文法ルールや辞書は、上述のように、上記分析データベース１８に予め準備されている。

【００６４】

なお、実施例では、一例として「茶筅」（http://chasen.naist.jp/hiki/ChaSen/）というフリーの形態素解析ソフトをコンピュータ１４に導入して利用した。

【００６５】
なお、文書が日本語の場合、実施例では、まず形態素を分割して抽出しその抽出した形態素に付いて品詞を付与するように、上記「茶筅」のようなツールを利用した。しかしながら、たとえば英語のような言語体系では形態素は既に分割されているので、形態素抽出処理は不要であるが、品詞を同定する必要があるので、このステップＳ５では、タギング（tagging：語の品詞を見分けること）処理をすることになる。
【００６６】
また、このステップＳ５で解析した形態素（群）および品詞情報は、テキストデータベース１６に蓄積される。
【００６７】
続くステップＳ７において、コンピュータ１４は、上述の品詞情報に基づいて、不要語として設定しておいた品詞の種類の形態素を取り除くための不要形態素除去処理を実行する。
【００６８】
つまり、形態素解析の際に、各形態素に付与される「品詞情報」に基づいて、当該形態素をキーワードの候補として採用するか否かを選定する。不要語とする形態素（特異語（キーワード）／共通語の候補）の品詞の種類は、形態素解析システムが出力する品詞体系と、ユーザの解析の意図によって異なる。不要形態素と認定する品詞の種類はユーザが任意で定められるものとする。発明者等が実際に解析を行なった実験では、「茶筅」を用いて分析した結果の、非自立や接尾の形を取らない名詞、動詞、副詞、形容詞以外を不要形態素とした。ただし、どのような品詞の形態素を不要語とするかという不要語除去規則もまた、分析データベース１８に予め設定しておけばよい。
【００６９】
ステップＳ７を実行した後には、たとえばテキストデータベース１６に蓄積されている当該コーパスの中に必要な１つ以上形態素が残っている。したがって、ステップＳ９－Ｓ１９の処理は、そのコーパスに除去されずに残っている各形態素毎に実行される。そのために、コンピュータ１４は、ステップＳ９において、適宜の規則で選定した順序に従って、処理すべき形態素を指定する。
【００７０】
次のステップＳ１１において、コンピュータ１４は、ステップＳ９で指定された形態素について、時間増加型ＴＦＩＤＦを求める。ここで、「ＴＦ」はTerm Frequency、つまり単位ドキュメント中にそのキーワード候補が出現する頻度(延べ数)（出現頻度）であり、時間のパラメータを考慮した「ＩＤＦ」は、Inversed Document Frequency（逆出現文書数）、つまり、他には出現していないという独自性を示す。したがって、「時間増加型ＴＦＩＤＦ」とは、「ＴＦ」×「ＩＤＦ」のことであり、Term Frequency Inversed Document Frequencyといい、ＴＦ＊ＩＤＦと表すこともあるが、ここでは、時間増加型ＴＦＩＤＦと表現する。時間増加型ＴＦＩＤＦは、当該形態素の出現率を示し、これは、一種の重み付け指標となる。
【００７１】
仮に、図５に示すように記事数が逐次変化する場合であっても、一般的な解析の場合には、最終的に一定数Ｎの単位ドキュメントが蓄積された後に行なうので、単位ドキュメントの総数Ｎは、図６（Ａ）に示すとおり一定数である。そのため、そのような一般のテキストデータを解析する際のＴＦＩＤＦのＤＦ（Document Frequency）、その形態素が出現する文書の数は、図７（Ａ）に示すように一定数となる。したがって、一般的な解析手法の場合のＴＦＩＤＦは図８（Ａ）のようになる。
【００７２】
これに対して、実施例のシステムで取り扱う１レコードは時間情報または順序情報２６（図２）を持っているため、各レコード（テキストデータ）は、時系列順または順序情報順に並べることができる。したがって、その際の時間増加型ＴＦＩＤＦのＤＦには、ｊの添え字（時間や順序の情報にもとづく添え字）が存在することになる。ここにいう「ｊ」

は、時系列順または順序情報順にレコードを並べた際の順番を表すことになる。

【００７３】

したがって、実施例の文書解析装置１０では、たとえば、ある記事ｄｊに対するＴＦＩＤＦを求める場合、最終的に収集された全件の記事に基づく単位ドキュメントの総数Ｎやそれに基づくＤＦを用いるのではなく、記事ｄｊが発行されるまでの時間に発信されていた記事の数に基づく時間を考慮したＮｊ（記事ｄｊが発信された時点までの記事の総数）や、ＤＦ（ｔｉ、ｄｊ）（記事ｄｊが発信された時点までの形態素ｔｉの出現文書数）を用いて、記事ｄｊが発信された時点で逐次ＴＦＩＤＦを計算する。この実施例の文書解析装置１０では、図４に示すようにそれが含む単位ドキュメント数が時系列順序にしたがって増加するコーパスを設定し、そのコーパスにおける各形態素のＴＦＩＤＦを計算することによって、時間的順序（順番）を有するテキストデータからその順序に従った特異語（キーワード）や共通語を抽出または検出する。

【００７４】

具体的には、通常のＴＦＩＤＦは次式（１）で、ここに定義する時間増加型ＴＦＩＤＦは次式（２）で計算される。

［数１］

$$TFIDF(ti,dj)=TF(ti,dj)*IDF(ti)$$

$$IDF(ti)=\log_{10}(N/DF(ti)) \quad \cdots\cdots\cdots(1)$$

［数２］

$$\text{時間増加型}TFIDF(ti,dj)=TF(ti,dj)*IDF(ti,dj)$$

$$IDF(ti,dj)=\log_{10}(Nj/DF(ti,dj)) \quad \cdots\cdots(2)$$

ここで、ｔｉはｉを識別子(ＩＤ)にもつ形態素である。つまり、ＴＦＩＤＦ（ｔｉ，ｄｊ）を算出する対象となるキーワード候補のことである。

【００７５】

ｄｊはｊ番目の単位ドキュメントを表わす。つまり、ＴＦＩＤＦ（ｔｉ，ｄｊ）および時間増加型ＴＦＩＤＦ（ｔｉ，ｄｊ）を算出する対象となるキーワード候補が含まれている文書のことである。ただし、文書の単位は、文章、記事、文など任意に設定可能であるが、実施例では、ウェブニュースの記事を文書単位とした。

【００７６】

ＴＦＩＤＦ（ｔｉ，ｄｊ）および時間増加型ＴＦＩＤＦ（ｔｉ、ｄｊ）は、ｊ番目の単位ドキュメントの形態素ｔｉ毎に算出される値である。

【００７７】

ＴＦ（ｔｉ、ｄｊ）は、ｊ番目の単位ドキュメントの形態素ｔｉごとに算出される値で、単位ドキュメントｄｊ中に形態素ｔｉが出現した回数（延べ数）である。

【００７８】

ＤＦ（ｔｉ、ｄｊ）は、１～ｊ番目の単位ドキュメント中に形態素ｔｉが出現した単位ドキュメント数である。

【００７９】

なお、上記Ｎｊは、単位ドキュメントｄｊが発生している際に出現している単位ドキュメント数であり、数度のＩＤが１から順序だって単位ドキュメントに付与されていれば実際には、Ｎの値はｊと同値になる。

【００８０】

たとえば図５に示すように、各記事（単位ドキュメント）ｄ１，ｄ２，ｄ３，…に出現する形態素ｔ１，ｔ２，ｔ３，…が変化する場合を想定する。この場合、単位ドキュメントの数Ｎｊをフィールドに持つテーブルが図６（Ｂ）に示すように表される。また、各単位ドキュメントのＤＦ（ｔｉ、ｄｊ）をフィールドに持つテーブルが図７（Ｂ）のように表され、Ｎｊの値によって形態素ｔｉを識別子にもった各単位ドキュメントの時間増加型ＴＦＩＤＦ（ｔｉ、ｄｊ）値をフィールドに持つテーブルが図８（Ｂ）のようになる。これらのテーブルは、いずれも、テキストデータベース１６に逐次蓄積される。

【００８１】

このようにして、ステップＳ１１で時間増加型ＴＦＩＤＦが計算された後、続くステップＳ１３において、コンピュータ１４は、時間増加型ＴＦＩＤＦの累計値Σ時間増加型ＴＦＩＤＦと、ＴＦの累計値ΣＴＦとをそのコーパスＣｔまでの実測値として計算する。なお、時間増加型ＴＦＩＤＦ（ｔｉ、ｄｊ）が図８（Ｂ）のようになり、ＤＦ（ｔｉ、ｄｊ）が図７（Ｂ）で表されることから、ＴＦ（ｔｉ、ｄｊ）も計算することができ、ΣＴＦについては、ＴＦ（ｔｉ、ｄｊ）を計算した後それの累計値として計算すればよい。ただし、時間増加型ＴＦＩＤＦについては、図８（Ｂ）のテーブルから累計値を計算すればよい。

【００８２】

続くステップＳ１５で、コンピュータ１４は、そのコーパスＣｔについて求めたＴＦ（ｔｉ、ｄｊ）の累積値ΣＴＦをＸとし、時間増加型ＴＦＩＤＦ（ｔｉ、ｄｊ）の累積値Σ時間増加型ＴＦＩＤＦをＹとして次式（２）への当て嵌めを行い、定数ａと定数ｂを求め、図９に示す回帰曲線を作成する。この回帰曲線は、次のコーパスＣｔ+Δｔでの残差分析のために、そのコーパスＣｔ+Δｔにおける時間増加型ＴＦＩＤＦを推定または予測するものとなる。つまり、そのコーパスＣｔまでのΣＴＦが横軸のようになるとき、もし、次のコーパスＣｔ+Δｔにおいても時間増加型ＴＦＩＤＦが同じ傾向を示すなら、次のコーパスＣｔ+Δｔでの時間増加型ＴＦＩＤＦは、この回帰曲線上にプロットされることになる。

［数３］

Ｙ＝ａＸｂ…… (3)

そして、コンピュータ１４は、ステップＳ１７において、先のステップＳ１３で計算した時間ｊでのコーパスＣｔにおける時間増加型ＴＦＩＤＦ（ｔｉ、ｄｊ）の累計値Σ時間増加型ＴＦＩＤＦと、前のコーパスＣｔ-Δｔについてステップ１５で求めた回帰曲線Ｙ＝ａＸｂによる推定値Ｙとの差（残差値）を求める（図１０）。残差値が大きいほど、正負のいずれに拘わらず、直前のコーパスＣｔ-Δｔで予測した同じ形態素ｔｉのΣ時間増加型ＴＦＩＤＦより離れている（乖離している）ことを、すなわち、直前のコーパスまでの常識から予測できなかったことを意味する。他方、Σ時間増加型ＴＦＩＤＦが正の残差値を示す形態素は、回帰曲線より上方にプロットされ、特異的または特徴的であることを意味する。Σ時間増加型ＴＦＩＤＦが負の残差値を示す形態素は、特異性は全くなく、逆の性質をもつありふれた形態素であるといえる。

【００８３】

図１０を参照して、Ｙ＝ａＸｂで示される回帰曲線に対して、形態素ｔｉのΣ時間増加型ＴＦＩＤＦがこの曲線の上方にプロットできた場合、この形態素ｔｉは正の残差値を持つことになる。正の残差値を持つということは、その形態素ｔｉがＣｔ-Δｔまでにあまり出現していないといえる。形態素ｔｉ+１のΣ時間増加型ＴＦＩＤＦは回帰曲線より下方にあり、したがって、この形態素ｔｉ+１はそれまでにも数多く出現した形態素であることを示している。

【００８４】

ステップＳ１７ではこのようにして各形態素毎にΣ時間増加型ＴＦＩＤＦの推定値または予測値と実測値との間で残差分析を行ない、各形態素の特徴値すなわち残差値を、たとえばデータベース１６のテキストデータテーブル２０（図２）にメタデータとして付加するなどして、逐次記憶する。

【００８５】

ステップＳ１９で最後の形態素について残差分析が終了したことを判断すると、コンピュータ１４は、次のステップＳ２１で、上述のようにデータベース１６に記憶した特徴値（残差値）に従って、特異語（キーワード）および一般語または共通語を選定する。たとえば、正の残差値が任意の上位数以上だった形態素を、そのコーパスを代表する特異語すなわちキーワードとして選定する。逆に、負の残差値が任意の下位数以下だった形態素は、一般語または共通語として選定する。一般語は構成したテキストデータベース（言語資料）全体を代表するキーワードに該当する。したがって、一般語を利用すれば、同じテー

マのテキストデータ（言語資料）を効率よく探し出せる。

【００８６】

続いて、コンピュータ１４は、最後のステップＳ２３で、ステップＳ２１で選定した特異語や共通語を図示しないディスプレイ上に表示する。

【００８７】

図１１の表示例では、表示画面の上側に正の残差値を持つ特異語が時間経過（横軸）とともにプロットされ、下側に負の残差値を持つ共通語がプロットされる。ただし、図１１では細部を描けないので、特異語として２つ「死亡」、「派遣」だけが明示されていて、共通語として「地震」、「新潟」という２つだけが明示されているが、各グラフ部分にそのグラフを構成する形態素（単語）が表示される、ということに留意されたい。この図１１のような表示例によれば、特異語と一般語が上下に別々に表示されているので、それらを一覧できるという利点がある。

【００８８】

表示例としては、図１２に示す表形式の表示も考えられる。図１２の表では、横軸に時間経過を示し、縦軸に時間区分ごとの特異語を上位適宜数表示するようにしている。

【００８９】

ただし、他の任意の表示形態が考えられることは勿論であり、図１１および図１２の表示例に限定されるものではない。

【００９０】

発明者等が実際に解析した実験では、２００４年新潟県中越地震（平成１６年１０月２３日１７：５６発生。Ｍ６．８。）について発行されたウェブニュースを用いた。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。

【００９１】

平成１６年（２００４年）１０月２３日以降に代表的なポータルサイト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行なった。収集した期間は、発災から翌年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２６２３件である。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。

【００９２】

６ヶ月間に収集した上記新潟県中越地震災害に関するウェブニュースのテキストデータを図２に示すテキストデータテーブル２０としてテキストデータベース１６(図１)に登録した。

【００９３】

その後、キーワード候補（形態素）を同定するために、ステップＳ５に従って形態素解析を実行してキーワードとして採用すべき言葉の単位を検討し、ステップＳ７に従って、ステップＳ５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いた。

【００９４】

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。

【００９５】
そこで、実施例では、一般に利用されている形態素解析によってキーワードとして抽出可能な単語を切り出すことにした。
【００９６】
形態素解析の結果の一例を示す：「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。
【００９７】
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。
【００９８】
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、１５２１１種類の形態素（合計６２３７６５の形態素）が得られた。
【００９９】
続いて不要語の除去を行なう。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持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（不要形態素）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。
【０１００】
こ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、この実施例で用いた形態素解析システムのもつ品詞体系が採用している品詞情報に基づいて、不要語とする品詞を決定する。
【０１０１】
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、１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、非自立や接尾の形をとるもの（「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。
【０１０２】
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と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。
【０１０３】
品詞情報をもとに不要語を除去した結果、形態素解析（ステップＳ５）で求められた１５２１１種類の形態素は、１４１０９種類にまで減少した（延べ５２１２４０の形態素）。１４１０９種類のうち、地震の発生から１～１０時間で１１２２種類の形態素（７２記事）、１０～１００時間で３５８１種類の形態素（４８１記事）、１００～１、０００時間で５６９１種類の形態素（１２３０記事）、１０００～４５２９時間で２７１６種類の

形態素（８４０記事）が出現した。

【０１０４】

次に、先に説明した式(１)に従って、ニュース記事から抽出したキーワード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を評価した。

【０１０５】

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。そこで、この実施例では、ステップＳ１１を実行して、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。

【０１０６】

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、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、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。

【０１０７】

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先に説明したＴＦＩＤＦがある。ここで、上述のように、ＴＦ（ｔｉ、 ｄｊ）がキーワードｔｉが記事ｄｊに出現した回数を示し、ＤＦ（ｔｉ）がキーワードｔｉが出現する文書数を示すとき、ＩＤＦ（ｔｉ）は、全文書数に対するキーワードｔｉが出現した文書数の比の逆数である。つまり、この実施例では、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との積をとった時間増加型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。

【０１０８】

そして、実施例では、ある記事ｄｊに対する時間増加型ＴＦＩＤＦを求める場合、最終的に収集された全２６２３件の記事に基づくＮやＤＦを用いることはせず、記事ｄｊ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Ｎｊ（記事ｄｊが発信された時点までの記事の総数）や、ＤＦ（ｔｉ、ｄｊ）（記事ｄｊが発信された時点\までの形態素ｔｉの出現文書数）を用いて、記事ｄｊが発信された時点で逐次ＴＦＩＤＦを計算することにする。これを時間増加型ＴＦＩＤＦと呼ぶ。

【０１０９】

時間の経過にともなって，増加するような言語資料体の例としては，危機・災害に関するものが挙げられる。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。実施例では、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基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ｄｊ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。前述のように、この時間の概念を考慮した指標を、時間増加型ＴＦＩＤＦと呼ぶ。

【０１１０】

ただし、単に時間増加型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。ある時点までの時間増加型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）ために時間増加型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時間増加型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。単純に時間増加型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。

【０１１１】

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。

【０１１２】

そこで、この実施例では、先に説明したように、残差分析(ステップＳ１７)を行なうことによって、ある時点と次の時点のコーパスの特性を比較するようにした。

【０１１３】

図１３に発災からそれぞれ１０時間（図１３（Ａ））、１００時間（図１３（Ｂ））、１０００時間（図１３（Ｃ））、４５００時間（図１３（Ｄ））までの形態素ごとのＴＦの累積値と時間増加型ＴＦＩＤＦの累積値の関係をプロットした。ＴＦの累積値と時間増加型ＴＦＩＤＦの累積値の間には、先の(２)式で表される強い関係があった。この(２)式の関数（線形関数）で両者の関係をみると、１０時間でＹ＝０．１６Ｘ＋３．１４（Ｒ２＝０．２４）、１００時間でＹ＝０．０７Ｘ＋１０．４７（Ｒ２＝０．１３）、Ｙ＝０．１１Ｘ＋１８．４６（Ｒ２＝０．１５）、Ｙ＝０．１５Ｘ＋２２．２７（Ｒ２＝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の累積値と時間増加型ＴＦＩＤＦ(の累計値の関係以外については、累乗関数でＲ２が０．９０～０．９９、線形関数でＲ２が０．１３～０．１７であり、ＴＦと時間増加型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。

【０１１４】

図１３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の累積値と時間増加型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、実際の時間増加型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時間増加型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、その逆で時間増加型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、実際の時間増加型ＴＦＩＤＦの累積値と、近似曲線に基づく推定値との差（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１４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。

【０１１５】

図１４の左側には、あるｔ－Δｔから単位時間幅Δｔ経過する際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(４)で表すことができる。

【０１１６】

【数４】

$$\bar{C}=Ct-\Delta t+C\Delta t\cdots\cdots\cdots\cdots(4)$$

ここで、$\bar{C}$は或る時間ｔにおけるコーパスであり、$Ct-\Delta t$は或る時間よりもΔｔだけ遡ったコーパスであり、$C\Delta t$は或る時間$t-\Delta t$からｔまでに増加したコーパスである。

【０１１７】

図１４（Ａ）に示すように、ＣΔｔ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たりしているような場合には、図１４の右上側に示したようにＴＦの累積値と時間増加型ＴＦＩＤＦの累積値の関係は、ｔ－Δｔ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、図１４（Ｂ）に示すように、ｔ－Δｔまでに出現しなかったようなキーワードがΔｔ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、図１４の右下側に示したようにＴＦの累積値と時間増加型ＴＦＩＤＦの累積値の関係を表す曲線の形状も大きく変化する。

【０１１８】

つまり、ある時点ｔでの時間増加型ＴＦＩＤＦの累積値と、ｔ－Δｔ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Δｔ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Δｔ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。

【０１１９】

このように、実施例では、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－Δｔのコーパスで構成されるＴＦと時間増加型ＴＦＩＤＦの累積値にもとづく関係式による時間増加型ＴＦＩＤＦの累積値の推定値とｔの時点での時間増加型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または特異語（残差値：正）、著しく低かったキーワードを一般語または共通語と呼ぶことにする（残差値：負）。

【０１２０】

図１に示す実施例の文書解析装置１０によれば、図３に示すフローチャートに示す次の手順に従って、コンピュータ１４によって、人の主観的な判断を用いず、時間増加型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。

【０１２１】

このようにして、図１に示す実施例の文書解析装置１０において、コンピュータ１４は、要するに、次のステップを実行する。

【０１２２】

1）時系列的に増加するテキストデータ（この場合では、ウェブニュース）のデータベースを構築する。

【０１２３】

2）テキストを形態素に分割し、品詞情報を付加する。

【０１２４】

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。

【０１２５】

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と時間情報に基づく時間増加型ＴＦＩＤＦを求める。

【０１２６】

5）ある時点ｔ－Δｔ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－Δｔまでのコーパスにおけるの累積値と時間増加型ＴＦＩＤＦの累積値の関係式を求め、それにもとづくｔの時点での時間増加型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Δｔに出現したキーワードの特徴量とする。

【０１２７】

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。

【０１２８】

実施例のシステムで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。

【０１２９】

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を丹念に採取することによって既に実現されている災害過程のモデルによれば、災害過程において時間は、１０時間、１００時間、１０００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。１００～１０００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。１０００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。

【０１３０】

この災害過程のモデルに基準とし、１～１０時間、１０～１００時間、１００～５００時間、５００～１０００時間、１０００～２０００時間、２０００～３０００時間、３０００～４５００時間の７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、３時間、８時間、８時間、２４時間、２４時間、２４時間に設定してキーワード検出を試みた。

【０１３１】

図１５‐図２１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。これらの図１５‐図２１のグラフは図１に示すコンピュータ１４のモニタ１５Ｂに表示される。図２２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。この図２２についてもモニタ１５Ｂで表示するようにしてもよい。

【０１３２】

図１５‐図２１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」が続いている。

【０１３３】

図１５‐図２１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。

［表１］

残差値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

１位	ボランティア	１４
２位	IC	１３
３位	断層	１１
４位	震度	9
	ダム	9
４位	児童	9
５位	レール	7
６位	電話	6
	起きる	6
６位	同市	6
	トンネル	6
	雨	6
	組合	6
	入居	6
７位	死亡	5

	羽田	5
7位	授業	5
	湖	5
	子ども	5
	判定	5
	除雪	5
8位	補助	4
	余震	4
8位	土砂崩れ	4
	今回	4
	可能	4
	ガル	4
	加速度	4
	星野	4
	村民	4
	優太	4
	排水	4
	回答	4
9位	道路	3
	自宅	3
	山	3
	義援金	3
	燕三条	3
	屋台	3
9位	選手	3
	雪下ろし	3
10位	防災	2
	派遣	2
	安否	2
	発生	2
	現在	2
	県内	2
	震源	2
	小国	2
	トイレ	2
	貴子	2
	保険	2
	優	2
	陛下	2
	大人	2
	紀宮	2
	補強	2
	募金	2
	業者	2
	旅館	2
	ペット	2
	移転	2

次に、図２２を参照して、検出されたキーワードの特徴量が時間の経過とともに変化していくのかについて考察する。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。

【０１３４】

図２２（Ａ）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２８６０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。

【０１３５】

図２２（Ｂ）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１００～１０００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、１０００時間以降で特徴量は負の値をとることになる。

【０１３６】

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。

【０１３７】

図２２（Ｃ）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１０００ 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１００～１０００時間、１０００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。

【０１３８】

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。

【０１３９】

また、図２２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に符合している。

【０１４０】

図２２で示さなかったキーワードの中でも、図１５‐図２１の上で高い特徴量を示しているものがある。１００～１０００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、特徴量が高まったとものと考えられる。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。

【０１４１】

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。

【０１４２】

「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１５‐図２１、図２２や表１に基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１５‐図２１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。

【０１４３】

キーワードの特徴量を用いた図１５‐図２１、図２２のような形で可視化（モニタ表示）を行なえば、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からデータや情報、知識や教訓を得ようとする際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。

【０１４４】

なお、上述の実施例では、設定時間毎にテキストコーパスを作成するようにした（Ｓ１，Ｓ３）。しかしながら、時系列的に増量するテキストデータをテキストデータベース１６に蓄積しておき、任意の時間幅Δｔ経過ごとにテキストブロックすなわちコーパスを画定するようにしてもよい。

【０１４５】

上で詳細に説明したように、この発明の解析手法は、語の出現分布について、任意の時点のコーパスＣｔとその時点からΔｔだけ遡った時点のコーパスＣｔ-Δｔを比較し、出現特性がｔ-Δｔとｔで大きく異なる特異的な語を特異語として抽出するものである。そのため、Δｔの中に、それまでに時系列的に増加してきたコーパスの語彙とは異なる語が出現した場合には、特異性を測定する特異値は高い値を示す。

【０１４６】

この発明の解析手法（アルゴリズム）では、特異値が高い値を示す場合には、以下の２つのパターンが想定される。１つは、ｔの時点で当該分野との関連が高く、当該分野に関連の高い語を多く含む文書（記事）がコーパスに加わった場合であり、１つは、ｔの時点で当該分野との関連があまり高くなく、当該分野との関連の低い語を含む文書がコーパスに加わった場合である。

【０１４７】

たとえば、発明者等が解析した２００７年新潟県中越沖地震のウェブニュースコーパスについていえば、特集記事群の中に、全国高校野球選手権の予選の結果を報じるニュースにおいて主な被災地である柏崎市の高校の試合結果が掲載されていたために、これがコーパスに加わった。この記事の中には、柏崎市の高校の試合結果だけでなく、その日行われた新潟県内の高校すべての試合の結果が掲載されていた。試合結果の中には、「二塁打○○本、三塁打○○本」という記述が多く含まれており、形態素「二塁打」および「三塁打」が著しく高い特異値を示した。

【０１４８】

このような後者の場合、時系列的に増加してきたコーパスの当該分野と関連の低い語に高い特異値を与えてしまうことになり、ときには、ユーザにニュースの把握を誤らせる結果を生じる可能性が否定できない。

【０１４９】

そこで、図２３以降で示すこの発明の他の実施例では、（１）Δｔにおいて出現文書数が１件である語（形態素）を除外するというフィルタリング１を施すことによって、極端に高い特異値を示す形態素を除外する方法、および／または（２）出現文書数と語（形態素）の出現頻との関係から、出現頻度が著しく高い形態素を除外するというフィルタリング２を施すことによって、極端に高い特異値を示す形態素を除外する方法を提案する。ただし、これらの方法を採用するかどうかはオプションとして、ユーザの選択に委ねることとした。

【０１５０】

さらに、この発明は、特異語（キーワード）を形態素の単位で解析を行い、可視化を行うものである。形態素を単位とする解析の欠点として、本来それぞれの形態素（特異語）がもつ文脈のもつ情報が失われ、高い特異性を示した語が何を表す言葉なのかの理解や解釈が難しいことにあった。そこで、以下の実施例では、注目すべき記事を表示することによって、文脈の情報を補完し、もって解析結果の理解や解釈を助長できる手法を提案する。

【０１５１】

図２３はこの発明の他の実施例の動作を示すフロー図である。この実施例は、上述のフィルタリングおよび注目記事表示のオプションを採り入れた実施例である。

【０１５２】

図２３において、ステップＳ１７までは、先の図３に示す実施例のステップＳ１‐Ｓ１７と同じであるため、ここでは重複説明を省略する。

【０１５３】

ただし、この実施例では、ユーザは、図２３の動作を開始する前に、図１に示す操作手段１５Ａを用いて、コンピュータ１４がモニタ１５Ｂに表示するＧＵＩ（図示せず）において、フィルタリングのオプションを採用するか、もし採用するならフィルタリング１およびフィルタリング２どちらを採用するのか、さらには注目記事表示のオプションを採用するのかどうか、を予め選択的に設定しておく。そして、このユーザ設定は、コンピュータ１４内のメモリ（図示せず）にたとえばフラグとして記憶しておく。フィルタリングオプションを選択しない場合、フィルタリングフラグが「０」として記憶され、フィルタリング１を選択した場合、フィルタリングフラグが「１」として記憶され、フィルタリング２を選択した場合、フィルタリングフラグが「２」として記憶される。そして、注注目記事表示オプションを選択したときには注目記事表示フラグが「１」として設定される。

【０１５４】

そして、ステップＳ１７まで実行した後、コンピュータ１４は、ステップＳ１８で、その語（形態素）の時間Δｔ内での出現頻度ＴＦ(Δｔ，ｔｉ)とその語（形態素）が時間Δｔ内で出現した文書（記事）数ＤＦ(Δｔ，ｔｉ)をコンピュータ１４のメモリにたとえば、図２４のような形式で、記憶する。ただし、これらの出現頻度ＴＦ(Δｔ，ｔｉ)および出現文書数ＤＦ(Δｔ，ｔｉ)は、先に説明ステップＳ１３において求められているものであり、このステップＳ１８ではそれらの数値を図２４に示すように記憶する。

【０１５５】

ただし、これらの出現頻度ＴＦ(Δｔ，ｔｉ)および出現文書数ＤＦ(Δｔ，ｔｉ)は、ユーザがフィルタリングのオプションを選択しなかった場合には、利用されることがない。この場合には、ステップＳ２０Ａで“ＹＥＳ”が判断され、ステップＳ２１で図３のステップＳ２１と同じ方法で、特異語および共通語（一般語）を選択し、ステップＳ２３に進み、このステップＳ２３で、たとえば図１５‐図２１に示すような、モニタ１５Ｂ上でのグラフ表示を行なう。

【０１５６】

フィルタリングオプションが設定されているときには、ステップＳ２０Ａで“ＮＯ”が判断されるため、続くステップＳ２０Ｂで、コンピュータ１４は、メモリ（図示せず）のフラグ領域を参照して、フィルタリングフラグが「１」かどうか判断する。このステップＳ２０Ｂで“ＹＥＳ”が判断されるということは、フィルタリング１がオプションとして選択されていることを意味し、“ＮＯ”が判断されるということは、フィルタリング２がオプションとして選択されていることを意味する。

【０１５７】

フィルタリング１がオプションとして選択されている場合、コンピュータ１４は、次のステップＳ２１Ａにおいて、フィルタ１で特異語・共通語を選択する。

【０１５８】

具体的には、ステップＳ１８で図２４でメモリに記憶しておいた各時間Δｔにおける各語の出現文書数ＤＦ(Δｔ，ｔｉ)のデータを参照して、ＤＦ(Δｔ，ｔｉ)＝１の場合の形態素ｔｉを除いて、ステップＳ２１と同じ手法で、特異語・共通語を選択する。

【０１５９】

フィルタリング２がオプションとして選択されている場合、コンピュータ１４は、次のステップＳ２１Ｂにおいて、フィルタ２で特異語・共通語を選択する。

【０１６０】

具体的には、ステップＳ１８でメモリに記憶しておいた出現文書数ＤＦ(Δｔ，ｔｉ)と、出現頻度ＴＦ(Δｔ，ｔｉ)を読み出し、まず、各時点で説明変数Ｘを各Δｔの出現文書数ＤＦ(Δｔ，ｔｉ)とし、目的変数Ｙを各Δｔの語の出現文書数ＤＦ(Δｔ，ｔｉ)としたＹ＝ａＸ＋ｂの回帰直線（図２５，図２６）を求める。同時にこの回帰曲線の９５％信頼限界を求める（図２５，図２６参照）。そして、メモリから読み出した現時点のΔｔの出現文書数ＤＦ(Δｔ，ｔｉ)および現時点のΔｔの語の出現頻度ＴＦ(Δｔ，ｔｉ)のデータと、その９５％信頼限界とを比較し、現時点のΔｔの語の出現頻度ＴＦ(Δｔ，ｔｉ)が正の９５％信頼限界を上回っていた場合、その語（形態素）ｔｉを除いて、ステップＳ２１

と同様に特異語・共通語を選択する。

【０１６１】

なお、図２５と図２６は同じ意味のグラフであるが、図２５が一般的表現であり、図２６は発明者の実験で出現した具体的例を示す。正負いずれにおいても、９５％信頼限界を超えた場合（正の場合は上回った場合）、その形態素を除外する。

【０１６２】

この実施例で、フィルタリングオプションを選択しなかった場合には、たとえば図２７に示すグラフ表示がステップＳ２３で行なわれるのに対し、フィルタリング１が選択された場合のステップＳ２３でのグラフ表示は図２８に示すようになる。両者を対比すれば、前者ではただ１つの記事に出現した形態素「二塁打」が高い特異値を持つ特異語として表示されるのに対し、後者ではその形態素「二塁打」はフィルタリング処理によって除去され、表示されない。その意味で、解析テーマとは無関係の特異語が表示されてしまうという不具合は解消されるものの、図２７と図２８とを対比すればわかるように、フィルタリング１では、その他の形態素も除外されるという傾向があるという点に注意しなければならない。

【０１６３】

フィルタリング２が選択された場合のステップＳ２３でのグラフ表示は図２９に示すようになる。フィルタリング２のオプションを実行した場合には、図２７と図２８とを対比すればわかるように、無関係の語「二塁打」は残ってしまうものの、その他の不要語が除去されるので、幾分見やすいグラフ表示となっている。

【０１６４】

ステップＳ２３で解析結果を可視化表示した後、ステップＳ２５で、コンピュータ１４は、メモリを参照して注目記事表示フラグが「１」かどうか判断する。“ＮＯ”の場合はそのまま終了するが、“ＹＥＳ”の場合には、ステップＳ２７で注目記事のモニタ１５Ｂでの表示ステップを実行する。

【０１６５】

具体的には、先のステップＳ１７で残差値を求めたときに、各時点で語ｔｉの特異値ＤＶｔｉのリストが作成されるので、Δｔにおける文書ごとに、その文書に含まれる特異語（特異値が高い上位１０の語）について、特異値の総和を求める（ＲＶ＝ΣＤＶｔｉ）。そして、この特異値の総和ＲＶの高い、たとえば上位３つの文書を「注目記事」として選定する。選択された「注目記事」について、少なくとも見出しと、本文に含まれる特異語（上位１０）を、たとえば表２に示すように表示する。

【０１６６】

上記特異値リストにリストアップした形態素ｔｉがどの文書に含まれていたかはたとえば図２のようなテキストデータテーブル２０を参照することによって特定できる。つまり、このステップＳ２７では、特異値総和ＲＶの高い形態素が含まれる文書番号（ＩＤ）の文書をデータテーブル２０から読み出すことによって、表２のような注目記事表示を実行する。

【０１６７】

【表２】

注目記事の表示例

１位:

RV=19.0, 活, 耐震

「日本原子力産業協会長が「原発安全性は確保」」

「日本経団連名誉会長で日本原子力産業協会会長の今井敬氏（新日鉄名誉会長）が十七日、松江市内で会見し、…

「中越沖地震で島根原発の消火体制を点検」

「新潟県中越沖地震で、東京電力柏崎刈羽原発の変圧器から火災が発生した問題で、…

３位:

RV=12.7, 電話

「＜中越沖地震＞２日目の夜　避難なお９０００人」

「新潟県中越沖地震は１７日、２日目の夜を迎え、柏崎市など７市町村にある１１１カ所の公民館などで被災者８９９５人が避難所生活、を強いられた。…

【０１６８】

表２では、特異値の総和ＲＶが「１９．０」の２つの語「活」および「耐震」を含む２つの記事と、特異値の総和ＲＶが「１２．７」の１つの語「電話」を含む１つの記事が、少なくとも見出しが、望ましくは本文も含めて表示される。それによって解析によって失われた形態素の文脈の情報を補完できるので、高い特異性を示した語が何を表す言葉なのかの理解や解釈が難しくなるのを回避することができる。

【０１６９】

ただし、上記実施例では、特異値総和ＲＶの高い上位３つの形態素についてそれらが含まれる「記事」すなわち単位ドキュメントを表示するようにしたが、何個の形態素について記事を表示するかは任意である。最上位の形態素についてだけそれを含む記事（見出し）を表示するようにしてもよく、上位１０個の形態素について記事や見出しを表示するようにしてもよい。

【０１７０】

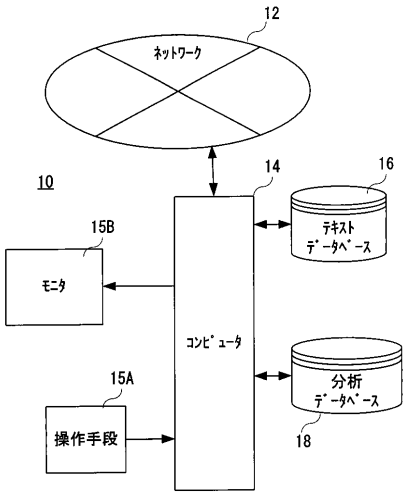
なお、選択した特異語や一般語を可視的に出力するために実施例ではそれらをモニタ上で表示するようにしたが、当然この表示に代えて、もしくはその表示に加えて、たとえばプリンタによって印刷出力することも可能である。

【０１７１】

なお、図１５‐図２１および図２７‐図２９においては、いくつかの描くべき特異語（キーワード）が省略されていることに留意されたい。理由は、図面内にできるだ余白を確保する必要があったためであり、したがって、スペースの狭い場所ではより多くの書くべき語が省略された。

【図１】

図1



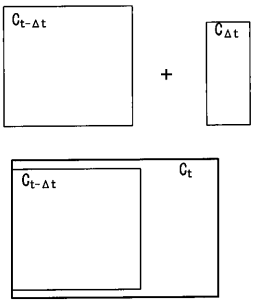
【図２】

図2

20 ID	22	26 時間（順序）情報	28 タイトル	24 本文（テキストデータ）
000000100057		2004.10.24 0:07	新潟地震 4人目の死亡	新潟県を中心とした地震で、同県によると、同県小千谷市の女性（70）がテーブル……
000000100058		2004.10.24 0:01	長野県内で1400戸停電	長野県などのまとめによると、同県内では、判断飯山市、栄村、野沢温泉村で約1400戸…
…		…	…	…

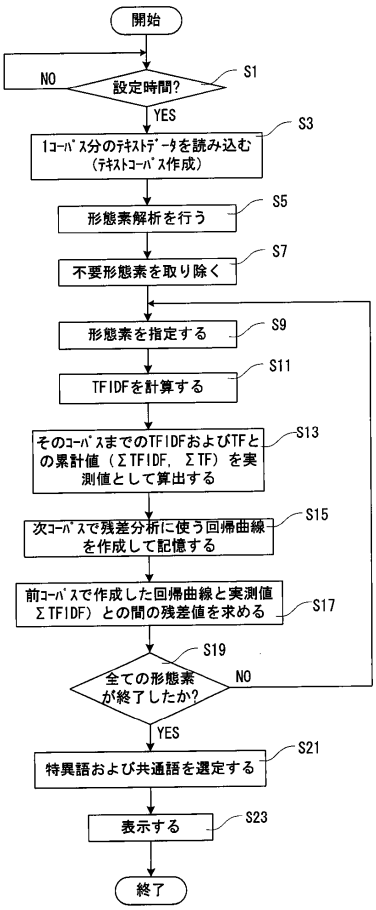
【図４】

図4



【図３】

図3



【図５】

図5

形態素	TF1	TF2	TF3	…
t1	2	0	1	
t2	1	3	0	
t3	0	1	0	
t4	0	0	1	
…				

【図６】

図6

(A)

	N
t1	3
t2	3
t3	3
t4	3
…	

(B)

形態素	N1	N2	N3	…
t1	1	2	3	
t2	1	2	3	
t3	1	2	3	
t4	1	2	3	
…				

【図７】

図7

(A)

	DF
t1	2
t2	2
t3	1
t4	1
…	

(B)

形態素	DF1	DF2	DF3	…
t1	1	1	2	
t2	1	2	2	
t3	0	1	1	
t4	0	0	1	
…				

【図８】

図8

(A)

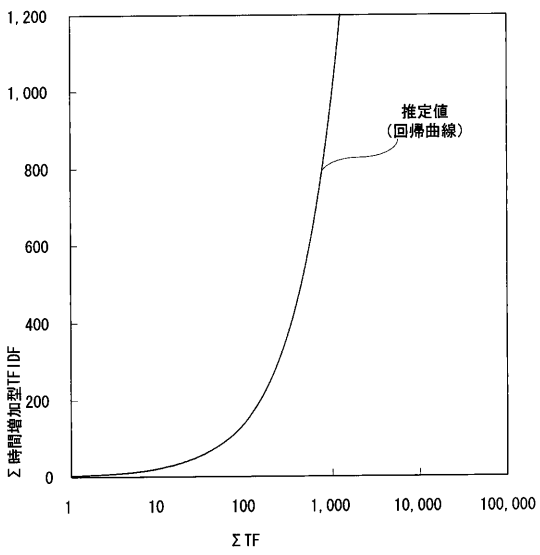
形態素	TFIDF1	TFIDF2	TFIDF3	…
t1	2*log3/2	0	1*log3/2	
t2	1*log3/1	3*log3/2	0	
t3	0	1*log3/1	0	
t4	0	0	1*log3/1	
…				

(B)

形態素	時間増加型 TFIDF1	時間増加型 TFIDF2	時間増加型 TFIDF3	…
t1	2*log1/1	0	1*log3/2	
t2	1*log1/1	3*log2/2	0	
t3	0	1*log2/1	0	
t4	0	0	1*log3/1	
…				

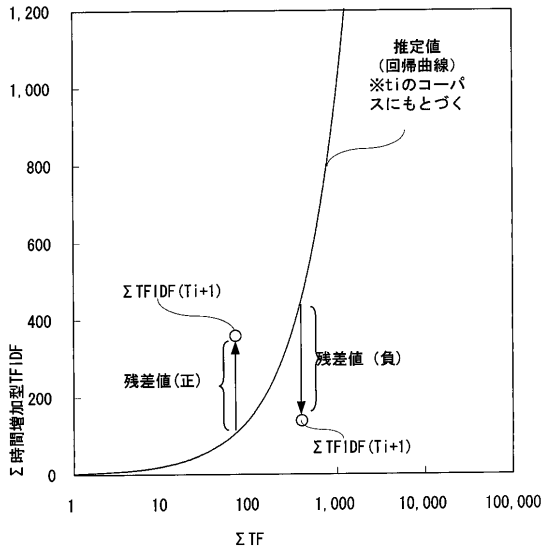
【図９】

図9



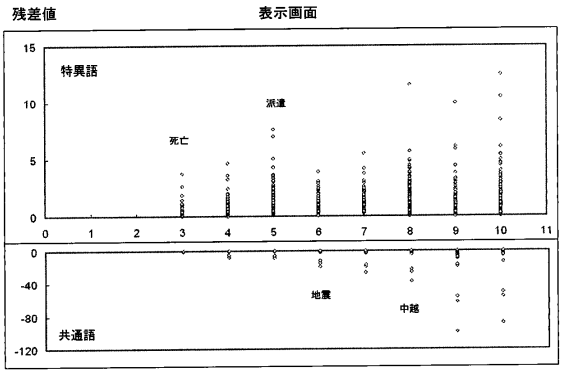
【図１０】

図10



【図１１】

図11



【図１２】

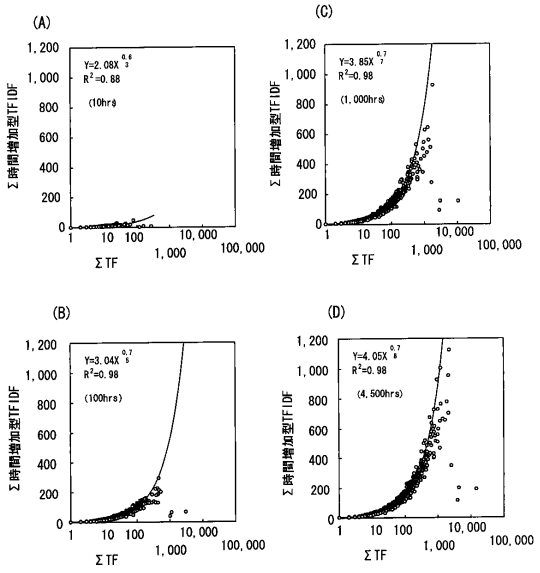
図12

特異語

Rank	時点T1の特異語	時点T2の特異語	時点T3の特異語	…
1				
2				
3				
4				
…				

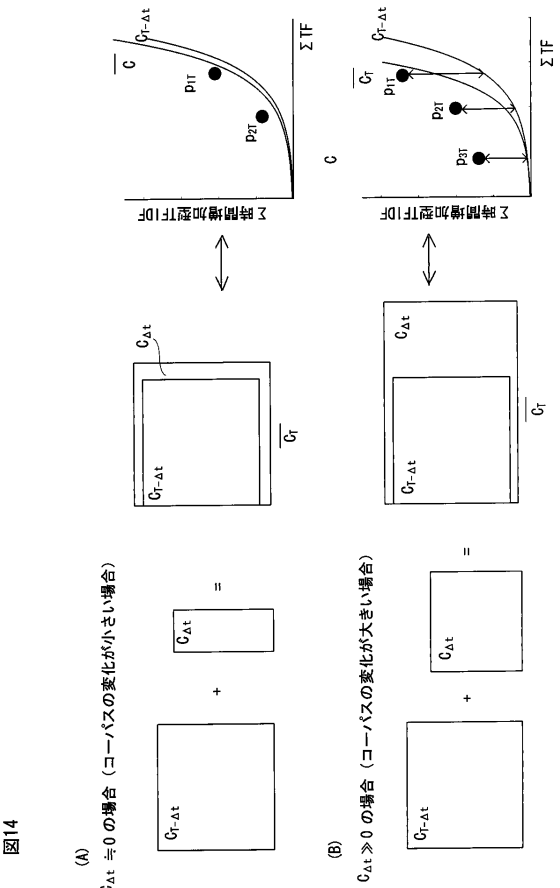
【図１３】

図13



【図１４】

図14



【図１５】

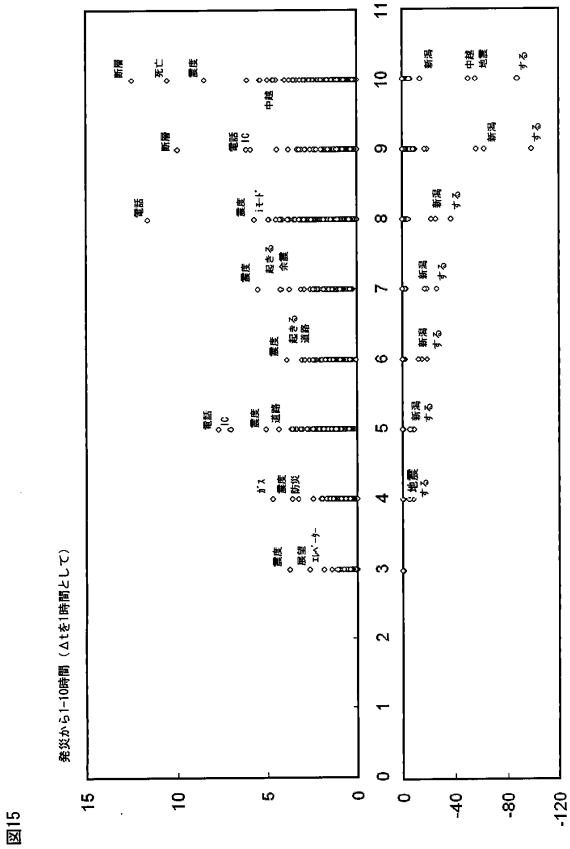


図15

【図１６】

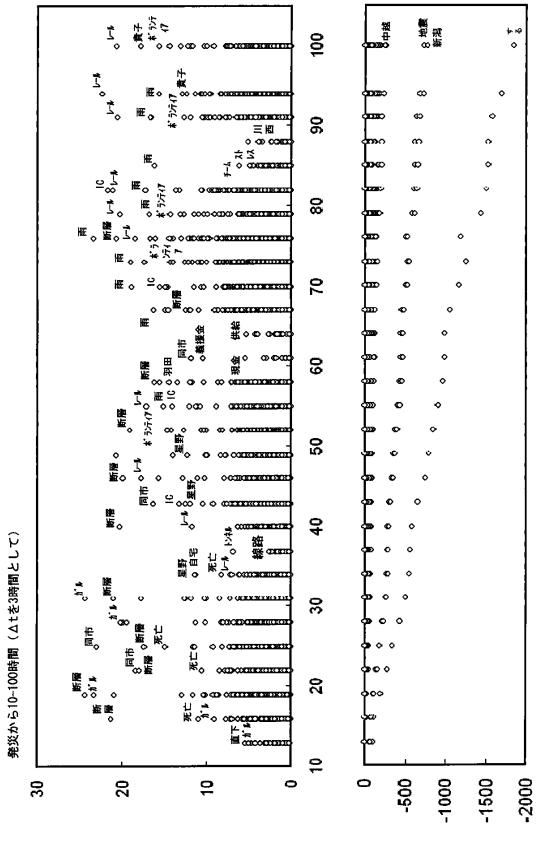


図16

【図１７】

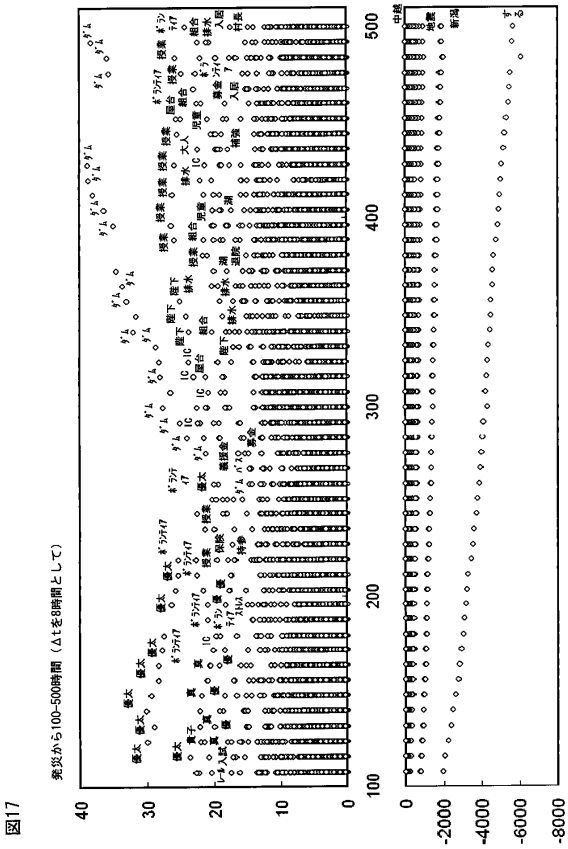


図17

【図１８】

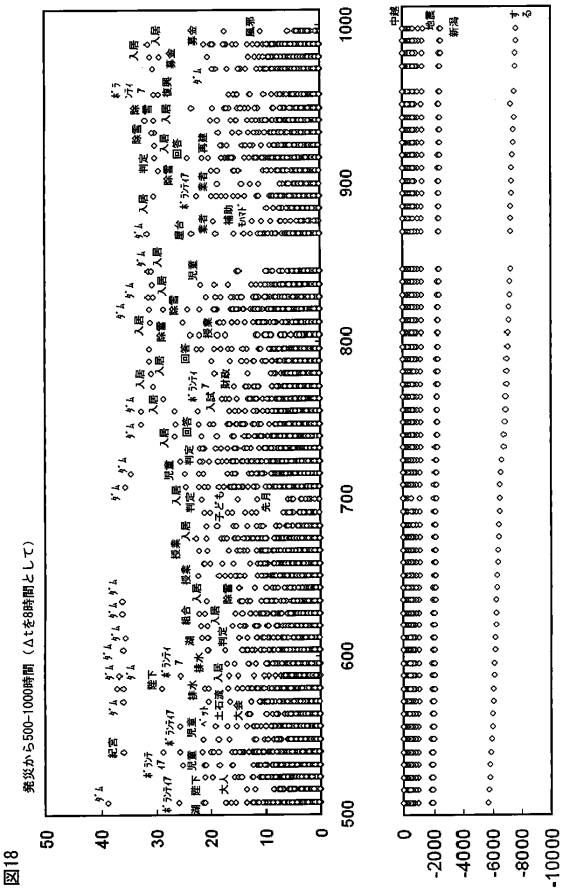


図18

【図１９】

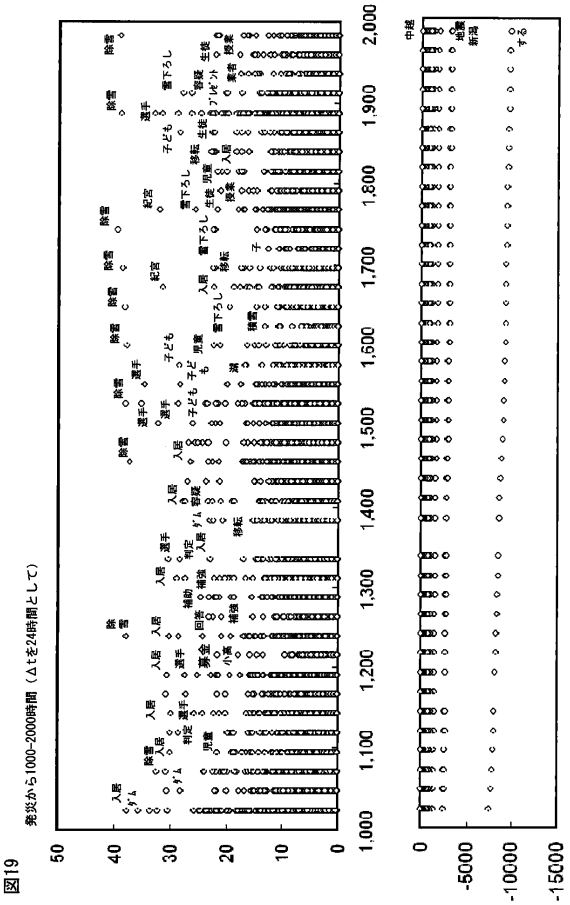


図19

【図２０】

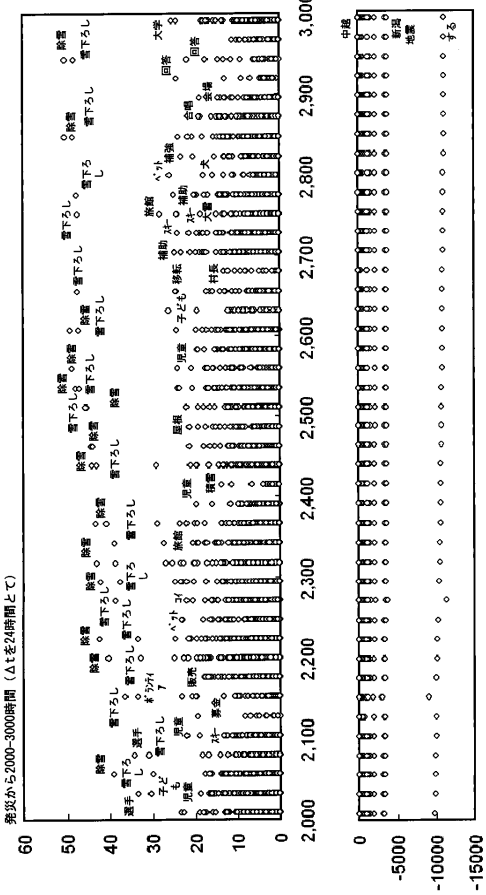


図20

【図２１】

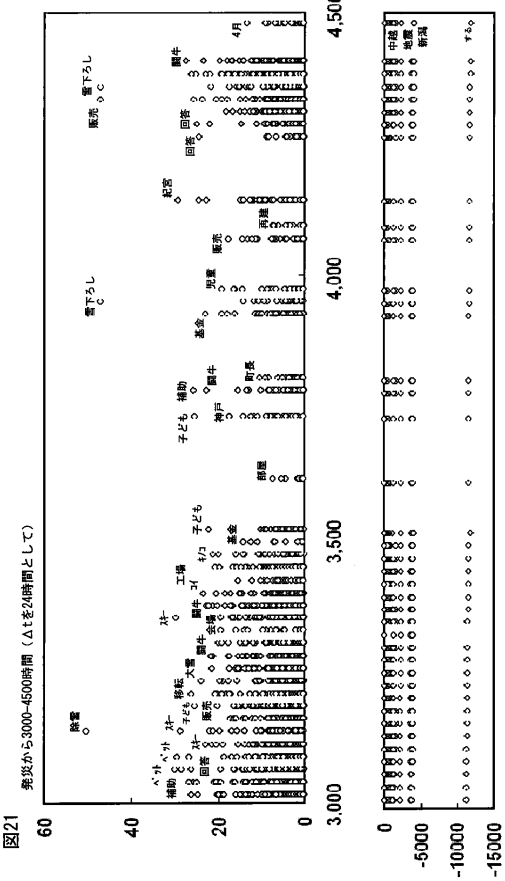


図21

【図２２】

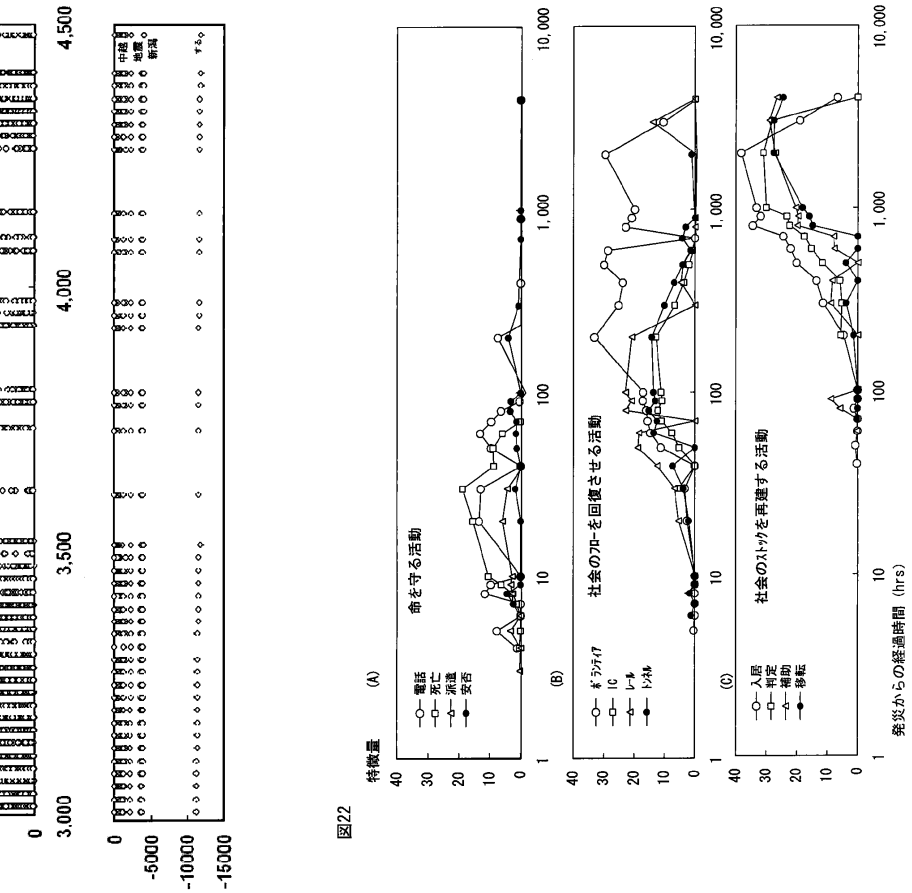
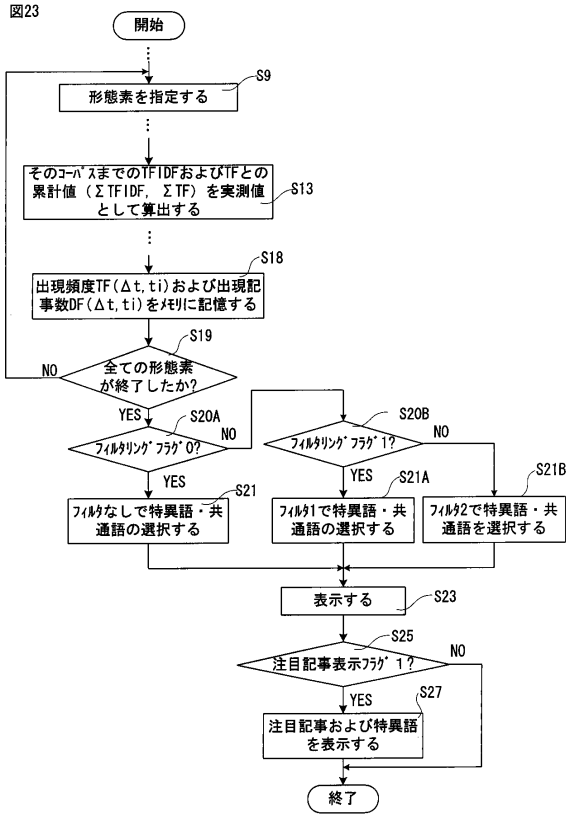


図22

【図２３】

図23



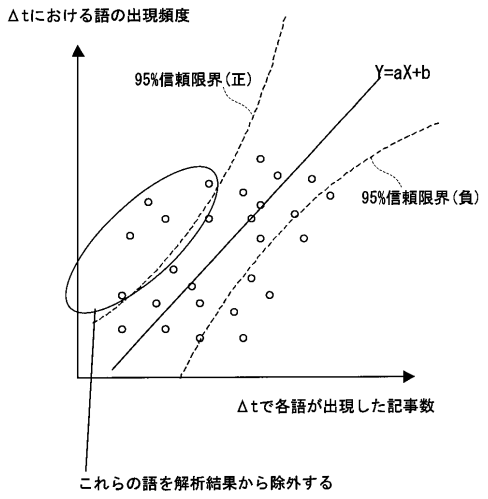
【図２４】

図24

形態素	出現頻度 TF	出現記事数 DF
t1	1	1
t2	2	1
t3	1	2
t4	4	3
…		

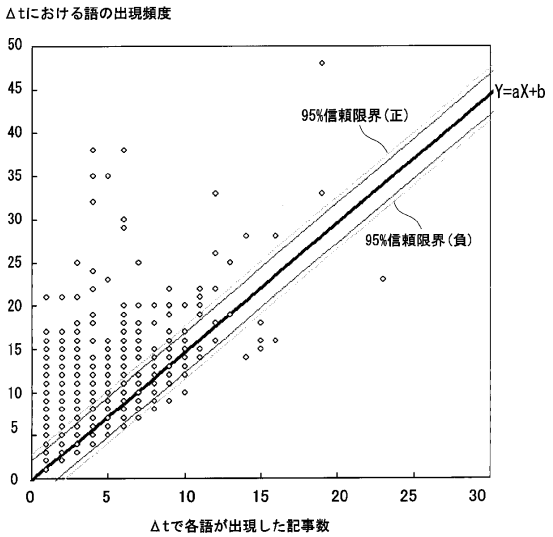
【図２５】

図25



【図２６】

図26



【図２７】

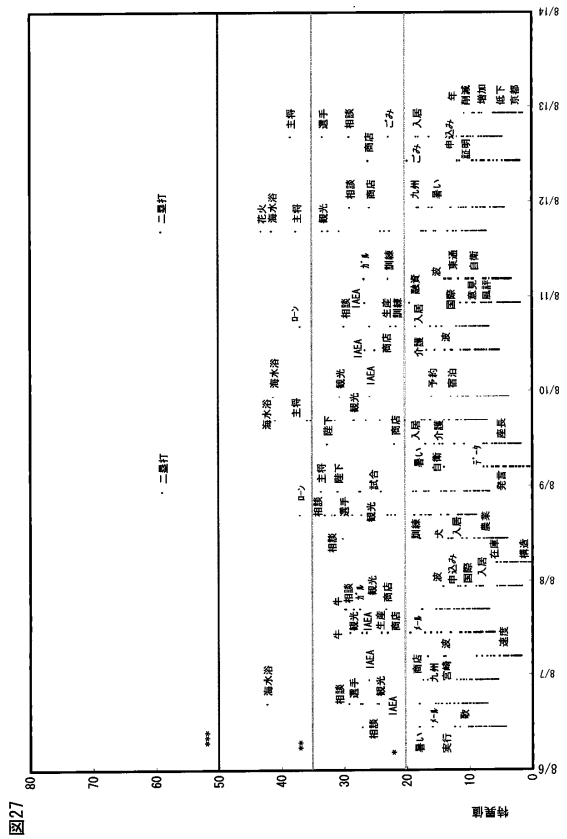


図27

【図２８】

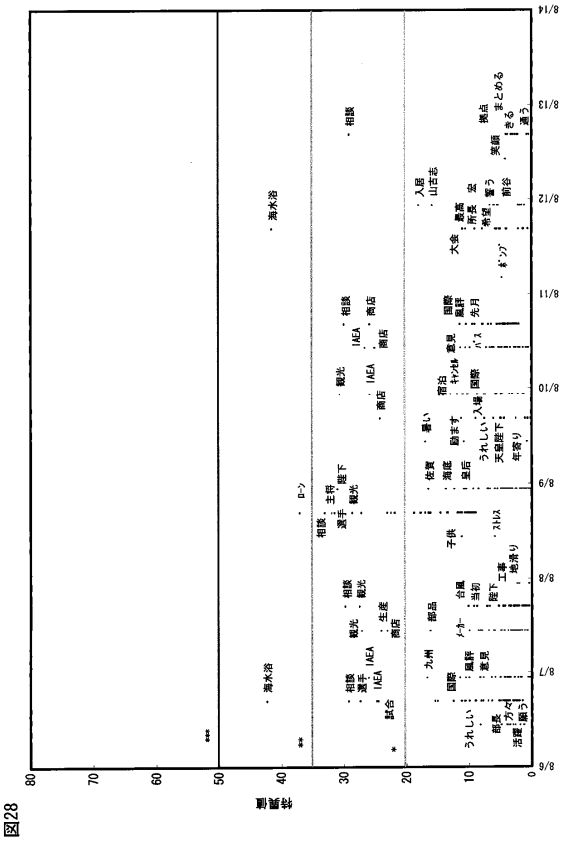


図28

【図２９】

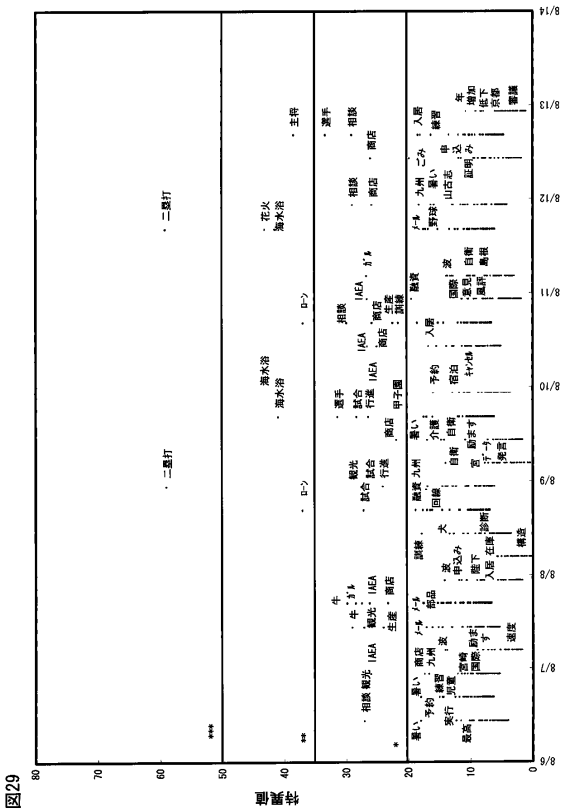


図29

フロントページの続き

(56)参考文献 特開２０００－１９４７４５（ＪＰ，Ａ）
特開２００５－３１６８９９（ＪＰ，Ａ）
特開平０８－０７７１７８（ＪＰ，Ａ）
特開２００３－１４１１３４（ＪＰ，Ａ）
特開２００５－３５２６１３（ＪＰ，Ａ）
大塚 真吾 他，検索語間の関連を考慮したＷｅｂ検索法の提案，電子情報通信学会技術研究報告〔データ工学〕，日本，社団法人電子情報通信学会，２００２年 7月１９日，Vol.102 No.209 (DE2002-75)，pp.85-90.
河合 英紀 他，WWW検索サービスにおけるトレンド語抽出，第５８回（平成１１年前期）全国大会 講演論文集，日本，社団法人情報処理学会，１９９９年 3月 9日，Vol.3 No.4T-3，pp.3-91～3-92.
那須川 哲哉 他，２テキストマイニング－膨大な文書データの自動分析による知識発見－，情報処理，日本，社団法人情報処理学会，１９９９年 4月１５日，第40巻 第4号，pp.358-364.

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/30
JSTPlus(JDreamII)