



【課題】必要十分な範囲内で抽出した追加学習データを分析処理することによりデータ集中を精度よく予測するデータ集中予測装置、その方法、及びそのプログラムを提供すること。【解決手段】データ入力手段 1 1 が取り込んだ時系列データのデータ構造をデータ記憶手段 2 1 を利用して分析すると共に後発するデータ集中を予測するデータ集中予測手段 3 1 は、各時系列データの入力時刻の直前の時刻を基準とした過去の一定期間内の時系列データに基づいて予め設定される変動許容範囲を逸脱する当該時系列データを後発するデータ集中の予測に必要な追加学習データとして連続的に抽出処理する学習データ抽出処理部 4 1 を備え、当該追加学習データに各種算出処理を施した加工情報に基づいて予測処理部 7 1 が将来のデータ集中にかかる予測値を算出する。

明 細 書

発明の名称：

データ集中予測装置、データ集中予測方法、及びそのプログラム

技術分野

[0001] 本発明は、将来のデータ集中の予測を目的として時系列データが内包する有意な情報を抽出すると共に、これに基づく学習及び予測処理を精度よく行うデータ集中予測装置、その方法、及びそのプログラムに関する。

背景技術

[0002] センシング技術および情報管理技術の発達に伴い、経時的に蓄積する時系列データから有用な知識（規則性等の情報）を抽出することが、機械学習（コンピュータプログラムを使って大量のデータから有用な知識を抽出する技術）やデータマイニング（統計学等のデータ解析の技法により有用な知識を取り出す技術）の分野における昨今の主要なテーマの一つとなっている。

[0003] ここで時系列データとは、地震の波形や津波の潮位変動といった自然現象に関するデータ、自動車や工場ラインに設置されたセンサから得られる各部品の状態に関するデータ、販売数量の変化に関するデータ、歩行者等の移動体による混雑についてのデータ、ツイッター（Twitter：登録商標）やブログ（Blog）などのウェブ（Web）上のソーシャルメディアへの経時的な投稿記事の内容や数、日常生活における電力使用量といった人間活動に関するデータなどを幅広く指す。また、これら各種の時系列データから抽出される有用な知識は、当該データの種類や利用目的によって異なる。

[0004] こうした有用な知識を抽出するための指標としては、例えば、与えられた時系列パターン（時系列データに現れるデータ構造のパターン）が正常であるか異常であるかを判定する基準や、ある時刻における観測値に基づいて未来の観測値を予測する数理モデルなどが挙げられる。このような指標を利用した有用な知識の抽出は、過去から現在に亘って蓄積された時系列データの背後に存在する本質的なデータ構造を学習することによって実現される。

[0005] ここで、判定基準や数理モデルのアルゴリズムは、経験的に得られた多様なデータに基づく統計的手法等を採用することが多いため、一般に、正常又は異常の判定や未来の観測値の予測にかかる精度は、学習データ（抽出し記憶させる有用な知識）の蓄積量がある程度多いほど安定する。また、時間の経過に伴い、これまで観測されなかった有意な時系列パターンが新たに出現した場合には、これを学習データとして追加処理することにより判定や予測の処理精度向上を図り得る。

[0006] 即ち、時系列データを対象とした学習装置による判定処理及び予測処理を精度よく行うためには、新たに蓄積される学習データを定期的に取り込むと共に、この取り込んだ新たな学習データをこれまでに蓄積された学習データと併せて再学習する機能が求められる。

[0007] その一方で、判定や予測を必要とする多くの応用場面では、再学習に適していないデータも混在し得るため、上記再学習機能を常時働かせる状況下では、不要な計算処理による時間の浪費が問題となる。また、有意な時系列パターンが含まれないデータを新たに学習データとして取り入れた場合には、それ以降の判定処理や予測処理の精度が低下する。さらに、こうした問題に対応して、新たに蓄積されたデータが学習データとして適切であるか否かについての評価をユーザがその都度行うようにした場合には、人的コストの増加や人為的ミスの発生といった新たな問題が生じる。

[0008] したがって、より好適な予測処理等を実現するためには、経時的に蓄積される時系列データに対して追加学習データとしての適切さを自動的に精度よく評価（判定）すると共に、当該評価により抽出された適切な追加学習データを有効に再学習する仕組みが必要となる。

[0009] こうした機械学習やデータマイニングの分野における関連技術としては、例えば、下記の技術内容（特許文献1乃至4、及び非特許文献1）が知られている。

[0010] 特許文献1には、販売量予測の高精度化と予測モデルの保守性向上を目的として、販売量実績の時系列データから多項式の回帰次数を計算し、実績値

が信頼限界の上下限値の範囲外となる際に当該月の販売量実績を異常値として抽出する手法が開示されている。この特許文献1では、当該手法により抽出した異常値を前月の実績値に基づいて補正することで、販売量予測の高精度化を期待している。

[0011] 特許文献2には、観測エリアの移動体の観測結果から予測モデルをリアルタイムに更新し、将来の混雑予測を行い、又、将来の移動体の到来について非定常ポアソン過程モデルを用いて仮発生させる方法が開示されている。この特許文献2では、混雑予測を要求された際に、要求時点での行列で表わされる移動体情報を初期値として予測モデルを更新する。

[0012] 特許文献3には、時系列データの過去の値から将来の値を予測する予測システムにおいて、時系列データの値が平常時の変動傾向から外れる変化点以前の値をノイズとして排除すると共に、それ以降の値を予測に用いるという技術内容が開示されている。特許文献4には、各種センサから入力される時系列データの中から異常発生時のデータを検出すると共に、当該検出データの前後に亘る予め設定された一定区間内の時系列データを抽出する技術内容が開示されている。

[0013] 非特許文献1には、学習データとしての適切さを自動的に評価するための能動学習と呼ばれる手法が記載されている。ここでは、既存の受動的学習アルゴリズムを用いて、リサンプリングされた複数の学習データ候補の中から相対的に適切なものを選択する技術内容が開示されている。

先行技術文献

特許文献

[0014] 特許文献1：特開平07-064965

特許文献2：特開2004-213098

特許文献3：特開2010-108283

特許文献4：特開2005-346655

非特許文献1：森下真一、宮野悟編「bit別冊 発見とデータマイニング」共立出版株式会社、2000年5月5日、pp. 64-72

発明の概要

発明が解決しようとする課題

- [0015] しかしながら、時系列データの異常値を補正し（特許文献1）、予測要求時点での時系列データを学習データとし（特許文献2）、又は、変化点以降の時系列データを学習データとする（特許文献3）手法や、これらを組み合わせた技術は、時系列データの変化そのものを予測する予測モデルの精度を向上させるものではない。また、特許文献4に開示された異常発生時の周辺データ抽出のための一定区間は予め設定しておく画一的な区間であり、当該区間を多様なデータの変動傾向に対応づけて柔軟に設定するという技術内容については何ら開示されていない。
- [0016] さらに、特許文献3には、時系列データが内包する有意な情報を抽出するという技術内容については何ら開示されていない。また、非特許文献1に記載された能動学習の手法において、与えられた学習データが絶対的に適切かどうかを判定するための技術内容については何ら開示されていない。
- [0017] （発明の目的）

本発明は、上記関連技術の有する不都合を改善し、特に、時系列データにおける後発するデータ集中を精度よく予測するデータ集中予測装置、その方法、及びそのプログラムの提供を、その目的とする。

課題を解決するための手段

- [0018] 上記目的を達成するために、本発明にかかるデータ集中予測装置では、複数のノードから発信されるデータに対応する属性データと共に受信し時系列データとして取り込むデータ入力手段と、この取り込んだ時系列データを記憶するデータ記憶手段と、この記憶した時系列データのデータ構造を分析し後発するデータ集中を予測するデータ集中予測手段と、を備えたデータ集中予測装置であって、データ集中予測手段は、データ入力手段が取り込む時系列データを予め設定された単位時刻ごとに経時的に一時記憶処理すると共に、この一時記憶処理データを利用して各時系列データの入力時刻の直前の時刻を基準とした過去の一定期間内の時系列データに基づいて予め設定される

変動許容範囲を逸脱する当該時系列データを後発するデータ集中の予測に必要な追加学習データとして連続的に抽出処理する学習データ抽出処理部を備え、この学習データ抽出処理部が、上記連続的に抽出した追加学習データをデータ記憶手段に一括して記憶処理する学習データ記憶処理機能を備えるという構成を採っている。

[0019] また、本発明にかかるデータ集中予測方法では、複数のノードから発信されるデータを対応する属性情報と共に受信し時系列データとして取り込むデータ入力手段と、この取り込んだ時系列データを記憶するデータ記憶手段と、この記憶した時系列データのデータ構造を分析し後発するデータ集中を予測するデータ集中予測手段と、を備えたデータ集中予測装置にあって、データ集中予測手段は、データ集中の予測用の時系列データを抽出し処理する学習データ抽出処理部を備えており、データ入力手段にて取り込まれた時系列データを予め設定された単位時刻ごとに経時的に一時記憶処理し、この一時記憶処理データを利用して各時系列データが各々の入力時刻の直前の時刻を基準とした過去の一定期間内の時系列データに基づいて設定される変動許容範囲を逸脱するものであるか否かを判定し、変動許容範囲を逸脱する旨判定した際に当該判定にかかる時系列データを後発するデータ集中の予測に必要な追加学習データとして連続的に抽出し、この連続的に抽出した追加学習データをデータ記憶手段に一括して記憶処理すると共に、これら一連の各工程内容を学習データ抽出処理部が順次実行し、データ記憶手段に一括記憶された追加学習データ及び既存の学習データにより特定される時系列データのデータ構造に基づいて、データ集中予測手段の予測処理部がデータ集中を予測するという構成を採っている。

[0020] さらに、本発明にかかるデータ集中予測プログラムでは、複数のノードから発信されるデータを対応する属性情報と共に受信し時系列データとして取り込むデータ入力手段と、この取り込んだ時系列データを記憶するデータ記憶手段と、この記憶した時系列データのデータ構造を分析し後発するデータ集中を予測するデータ集中予測手段と、を備えたデータ集中予測装置にあっ

て、データ入力手段にて取り込まれた時系列データを予め設定された単位時刻ごとに経時的に一時記憶処理するデータ一時記憶処理機能、この一時記憶処理データを利用して各時系列データが各々の入力時刻の直前の時刻を基準とした過去の一定期間内の時系列データに基づいて設定される変動許容範囲を逸脱するものであるか否かを判定する変動許容判定機能、変動許容範囲を逸脱する旨判定した際に、当該判定にかかる時系列データを後発するデータ集中の予測に必要な追加学習データとして連続的に抽出する学習データ抽出機能、およびこの連続的に抽出した追加学習データをデータ記憶手段に一括して記憶処理する学習データ記憶処理機能、この一括記憶された追加学習データ及び既存の学習データにより特定される時系列データのデータ構造に基づいてデータ集中を予測する予測処理機能を設け、これら各情報処理機能を、データ集中予測手段が備えているコンピュータに実現させるようにする、という構成を採っている。

発明の効果

[0021] 本発明は、上述したように、取り込んだ時系列データの中から有意な追加学習データを連続的に抽出するに際して上記学習データ抽出処理部が有効に機能し、この抽出された追加学習データ及び既存の学習データに基づく分析処理を経て予測処理部が後発するデータ集中を予測するという構成を採用したため、これによると、特に、時系列データにおける後発するデータ集中をリアルタイムで精度よく予測することができるという優れたデータ集中予測装置、その方法、及びそのプログラムの提供が可能となる。

図面の簡単な説明

[0022] [図1]本発明の第1実施形態にかかるデータ集中予測装置の構成を示すブロック図である。

[図2]図1に開示したデータ集中予測装置による学習データの抽出処理及びこの学習データに基づく算出処理の動作を示すフローチャートである。

[図3]図1に開示したデータ集中予測装置による将来のデータ集中の予測動作を示すフローチャートである。

[図4]図 1 に開示したデータ集中予測装置による変動許容範囲の算出の際に基準とする直近過去 Δ 時刻の一例を示す説明図である。

[図5]図 1 に開示したデータ集中予測装置によって抽出する学習データ及び有効学習期間の一例を示す説明図である。

[図6]本発明の第 2 実施形態にかかるデータ集中予測装置の構成を示すブロック図である。

[図7]図 6 に開示したデータ集中予測装置による学習データの抽出処理及びこの学習データに基づく算出処理の動作を示すフローチャートである。

[図8]図 6 に開示したデータ集中予測装置による将来の投稿件数の予測処理動作を示すフローチャートである。

[図9]ノードとグループの関係に基づく影響力の算出の一例を示す説明図である。

発明を実施するための形態

[0023] 〔第 1 実施形態〕

本発明の第 1 実施形態にかかるデータ集中予測装置を、図 1 乃至図 5 に基づいて説明する。

[0024] 〔全体的構成〕

本第 1 実施形態において、図 1 に示す符号 8 1 は、外部から受信した時系列データの特徴点を抽出すると共にこれに基づいて後続するデータ集中を予測するデータ集中予測装置を示す。

このデータ集中予測装置 8 1 は、複数のノードから発信されるデータに対応する属性情報と共に受信し時系列データとして取り込むデータ入力手段 1 1 と、この取り込んだ時系列データを記憶するデータ記憶手段 2 1 と、この記憶した時系列データのデータ構造を分析し後発するデータ集中を予測するデータ集中予測手段 3 1 と、を備えている。

[0025] ここで、ノード（発信元）とは、ネットワークを構成する一つ一つの要素のことであり、コンピュータネットワークでは、サーバ、クライアント、ハブ、ルータ、アクセスポイント等を指し、センサネットワークでは、センサ

端末を指す。また、データ集中の予測をするデータとしては、各種センサが感知する情報、ブログやツイッターにおける投稿件数、地震の揺れの強さを示す応答速度、日々の電力使用量等の多様なデータを想定する。

[0026] データ集中予測手段 3 1 は、データ入力手段 1 1 が取り込む時系列データを予め設定された単位時刻（単位集計時間）ごとに経時的に一時記憶処理すると共に、この一時記憶処理データを利用して各時系列データの入力時刻の直前の時刻を基準とした過去の一定期間内の時系列データに基づいて予め設定される変動許容範囲を逸脱する当該時系列データを後発するデータ集中の予測に必要な追加学習データとして連続的に抽出処理する学習データ抽出処理部 4 1 を備えている。

[0027] また、データ記憶手段 2 1 は、学習データ抽出処理部 4 1 が連続的に抽出した追加学習データを記憶する学習データ記憶部 2 1 A と、この記憶された追加学習データに基づいてデータ集中予測手段 3 1 が行う各種算出処理の結果情報を記憶する学習処理情報保存部 2 1 B と、から成る。

[0028] 学習データ抽出処理部 4 1 は、連続的に抽出した直近の過去数時刻の変動傾向と比較して特徴的な振る舞いを示す追加学習データ（学習に必要である場合にはその近辺の時刻の時系列データも含む）を学習データ記憶部 2 1 A に一括して記憶処理する学習データ記憶処理機能 4 1 A を有している。

[0029] 即ち、変動許容範囲を逸脱している旨の判定（具体的には後述する）が連続している間は学習データ記憶部 2 1 A への記憶処理を行わず、変動許容範囲を逸脱している旨の判定が途絶えた際にまとめて一回で記憶処理するように構成されている。これにより、学習データ抽出処理部 4 1 は、有効な追加学習データが連続的に出現する期間を有効学習期間として抽出することができる。

[0030] データ集中予測手段 3 1 は、時系列データが内包する予測属性情報（データ集中の予測にかかる属性のデータ）を単位時刻に対応づけて集計データとして集計する情報集計部 6 1 を備えている。ここで、予測属性情報とは、時系列データに含まれる属性の中で将来のデータ集中を予測したい属性にかか

る情報である。

[0031] この情報集計部 6 1 は、学習データ記憶部 2 1 A に一括記憶された追加学習データが内包する予測属性情報を単位時刻に対応づけて学習集計データとして集計する学習データ集計機能 6 1 A と、外部入力により予め設定された時間間隔（予測間隔）で発せられる予測要求に応じてデータ入力手段 1 1 が取り込んだ時系列データに含まれる予測属性情報を単位時刻に対応づけて予測集計データとして集計する予測データ集計機能 6 1 B を有している。また、予め外部入力により設定された予測間隔にかかる情報はデータ記憶手段 2 1 に記憶されているものとする。本第 1 実施形態では、この予測間隔にかかる情報を予測処理部 7 1 がデータ記憶手段 2 1 から取得すると共に当該情報に沿ってデータ入力手段 1 1 に対し予測要求を行うように構成されている。

[0032] また、データ集中予測手段 3 1 は、データ集中にかかる予測値に対する各ノードの影響力を示す影響力データを学習データ集計機能 6 1 A から出力される学習集計データとの関係で算出すると共に、この影響力データ及び当該算出に用いた学習集計データを学習処理情報保存部 2 1 B に保存処理する学習処理部 5 1 と、前述した情報集計部 6 1 が予測データ集計機能 6 1 B にて集計した予測集計データと学習処理部 5 1 が算出した影響力データとに基づいて予測属性情報の予測値を算出処理する予測処理部 7 1 と、を備えている。この予測処理部 7 1 は、予測結果を装置外部に送信する予測結果出力機能 7 1 A を有している。更に、算出した予測値を予測処理部 7 1 がデータ記憶手段 2 1 に記憶処理するような構成としてもよい。

[0033] 学習処理部 5 1 は、更に、リアルタイムでの影響力データの算出に際して学習処理情報保存部 2 1 B に予め影響力データ及び学習集計データが保存されていた場合には、リアルタイムにて算出した影響力データ及び当該算出に用いた学習集計データにより学習処理情報保存部 2 1 B 内の保存情報を更新するデータ更新処理機能 5 1 A と、予め学習処理情報保存部 2 1 B に保存された学習集計データとリアルタイムにて情報集計部 6 1 から取得する学習集計データとを結合処理すると共にこれを使用して影響力データを算出する再

学習処理機能 51B と、を有している。

[0034] 学習データ抽出処理部 41 は、上述したように、追加学習データを抽出するに際して、データ入力手段 11 から取り込んだ時系列データを単位時刻ごとに区分して揮発性メモリ（図示せず）に一時記憶し（データ一時記憶処理機能）、これに基づいてリアルタイムの時刻の時系列データに含まれる予測属性情報に対する変動許容範囲を算出するように構成されている。

[0035] 本第 1 実施形態では、上記変動許容範囲を、各々の入力時刻から見た過去の一定期間内の時系列データが内包する予測属性情報の平均値と分散に基づいて学習データ抽出処理部 41 が算出し設定するように構成されている。ここで、図 4 に基づいて学習データ抽出処理部 41 による前述した変動許容範囲の算出手法を説明する。

[0036] 図 4 に示すように、時間を示す横軸は単位時刻（単位集計時間）ごとに区分され、例えば、 $T \sim T+1$ 間の時刻が時刻 $T+1$ である。

ここで、現時点の属する時刻（現在時刻）を時刻 T とした場合、学習データ抽出処理部 41 は、そこから遡って連続する合計 U の時刻（直近過去 U 時刻： $T-U+1, \dots, T$ ）を過去の一定期間として特定し、この直近過去 U 時刻内に発信された時系列データに含まれる予測属性情報の平均値と分散に基づいて「平均値 $\pm \alpha \times$ 標準偏差（分散の正の平方根）」により定まる変動許容範囲を算出する（ α ：逸脱に対する感応度）。

[0037] ここで、外部入力されるパラメータである感応度 α は、上記式の重要な要素であることから、追加学習データの抽出に大きく影響するものであり、このため、当該追加学習データに基づいて行うデータ集中の予測の精度にも大きな影響を及ぼすものである。そこで本第 1 実施形態では、仮に設定した幾つかの α の値に基づいて仮予測値を別途算出し、これら複数の仮予測値の内でも予測精度が良かったものに対応する α の値を採用するようにした。ここでの仮予測値は、過去に記憶処理した時系列データを用いて算出するようにした。

[0038] 学習データ抽出処理部 41 は、リアルタイムの時刻での予測属性情報が、

上記により算出した変動許容範囲を逸脱しているか否かを判定するように構成されている（変動許容判定機能）。即ち、リアルタイムの時刻での予測属性情報が、変動許容範囲内にあれば逸脱していない旨判定し、変動許容範囲から外れていれば逸脱している旨判定するように構成されている。

[0039] この変動許容範囲を逸脱している旨判定した学習データ抽出処理部41は、当該判定にかかる時系列データを追加学習データとして抽出するように構成されている。ここでの抽出とは、学習データ抽出処理部41が、当該判定にかかる時系列データ（学習に必要である場合にはその近辺の時刻の時系列データも含む）とそれ以外の時系列データとを区別できる状態にすること指す。

[0040] また、学習データ抽出処理部41は、上記「平均値 $\pm\alpha\times$ 標準偏差」という式の構成から、突発的な機械トラブル等に起因する突飛なデータが出現しない限り当該逸脱している旨の判定を連続的に行う。このため、当該判定にかかる時系列データは、学習データ抽出処理部41により連続的に抽出される。

[0041] 即ち、ある時刻 $T+1$ から連続する合計 B の時刻（連続する B 時刻） $T+1, \dots, T+B$ のすべての時刻において逸脱している旨判定した場合において、時刻 T までの時刻及び時刻 $T+B+1$ 以降の時刻においては逸脱していない旨判定した状況下では、時刻 $T+B+1$ の判定が終了した際に、学習データ抽出処理部41が、当該連続する B 時刻に対応する時系列データをまとめて一回で出力するように構成されている（上記例によれば、単位時刻ごとに B 回に分けて出力するものではない）。

[0042] （有効学習期間）

ここで、学習データ抽出処理部41が抽出する有効学習期間を図5に例示したグラフ（平均値と分散による有効学習期間の抽出イメージ）に基づいて説明する。

[0043] この図5に示すグラフは、横軸に時間（時刻）をとり、縦軸に観測値としての予測属性情報（データ集中の予測にかかる属性のデータ）をとったものである。観測値としては、地震の揺れの強さを示す応答速度、各種センサに

よる測定情報、ツイッターやブログにおける投稿件数、日々の電力使用量等の多様なデータを想定する。

[0044] 図5に示すように、太線で挟まれた区間「R(1)、R(2)、R(3)」は、時刻ごとの観測値が変動許容範囲を逸脱している時刻（直近過去数時刻の平均値から大きく変化している時刻）の連続期間である有効学習期間を示す。

[0045] 上述した学習データ抽出処理部41による抽出の手法によれば、有効学習期間が区画する範囲は、図5に示すR(1)、R(2)、R(3)のように波形の変動状況に応じて柔軟に変化する。このため、学習データ抽出処理部41は、後発するデータ集中の予測用の追加学習データを必要十分な範囲内において精度よく抽出することができる。

[0046] また、この有効学習期間は、過去の一定期間の観測値にかかる前述した「 $\text{平均値} \pm \alpha \times \text{標準偏差}$ 」により定まる変動許容範囲に応じて変化するものであり、具体的には、「どの程度の過去の時刻を直近の時刻とみなすか（過去の一定期間の範囲）」と「逸脱に対する感応度(α)の大きさ」によって変化する。即ち、 U の値と α の値を変化させることにより、学習データ抽出処理部41は、適宜必要とされる追加学習データを柔軟な有効学習期間内全域に亘って抽出することができる。

[0047] （前処理）

ここで、前述した学習処理部51が実行する前処理について説明する。

この前処理とは、新たな追加学習データが学習データ抽出処理部41にて抽出・追加された際、効率的な再学習（影響力データの算出等）が実現できるように算出処理等で得た元データに対して加工を施したもの（前処理データ）を保存しておくことである。

[0048] 例えば、代表的な数理モデルである回帰分析モデルの学習では、学習結果として回帰係数を得ることができる。この学習結果を取得する過程においては、元データを加工した目的変数(ベクトル)および説明変数(行列)の値を保存しておく構成を採用することにより、それ以降の学習段階（学習フェーズ）では、既存の前処理データを用いての効率的な再学習が可能となる。

[0049] 本第1実施形態にかかるデータ集中予測装置81では、図1に示すように、学習処理部51が影響力データを算出した際に、この影響力データ及び当該算出に用いた学習集計データを前処理データとして学習処理情報保存部21Bに保存処理するように構成されている。また、当該前処理データが学習処理情報保存部21B内に予め保存されていた場合には、データ更新処理機能51Aにて前処理データとしての当該保存情報を更新するように構成されている。こうして保存・更新された前処理データは、上述したように、再学習処理機能51Bによる影響力データの算出に際して利用される。

[0050] このようにすれば、新たな追加学習データが抽出・追加された際に、それが既存の前処理データに対応するものであれば当該前処理の計算を省くことができるため、これにより、再学習に要する計算時間の短縮が可能となる。

[0051] (動作説明)

次に、図1に示すデータ集中予測装置81の動作制御を図2又は図3に示すフローチャートに基づいて説明する。

[0052] (学習処理)

まず、図2に基づいて時系列データの学習処理を説明する。

データ入力手段11は、装置外部から時系列データを入力すると共に、これを学習データ抽出処理部41に送り込む(図2:S201)。

[0053] 続いて、学習データ抽出処理部41は、データ入力手段11から受け取った時系列データが内包する予測属性情報の直近での変動傾向に応じた変動許容範囲を算出すると共に、リアルタイムの時刻での時系列データに含まれる予測属性情報が当該変動許容範囲を逸脱しているか否かを判定する(図2:S202)。

[0054] 例えば、現在時刻Tの一時刻先であるT+1時刻の観測値(予測属性情報)が得られた際に、学習データ抽出処理部41は、その観測値が、直近過去U時刻における「 $\text{平均値} \pm \alpha \times \text{標準偏差}$ 」により特定される範囲外にあれば、変動許容範囲を逸脱している旨(直近過去U時刻の傾向から外れている旨)判定すると共に(図2:S202/はい)、当該判定にかかる時系列デー

タを追加学習データとして抽出し、後続する時系列データの判定処理に移行する（図2：S203）。このとき、学習に必要である場合には、学習データ抽出処理部41は、当該判定にかかる時刻近辺の時系列データも併せて抽出する（図2：S203）。

[0055] 一方で、 $T+1$ 時刻の観測値が、直近過去 U 時刻における「平均値 $\pm \alpha \times$ 標準偏差」により定まる範囲内であれば、学習データ抽出処理部41は、変動許容範囲を逸脱していない旨（直近過去 U 時刻の傾向に沿っている旨）判定し、当該判定にかかる時系列データを抽出せずに後続する時系列データの判定処理に移行する（図2：S202／いいえ）。

[0056] 上記抽出処理までの各工程内容（図2：S201～S203）が経時的に繰り返される本第1実施形態において（ $T+1, T+2, T+3, \dots$ ）、変動許容範囲を逸脱する観測値は「平均値 $\pm \alpha \times$ 標準偏差」という式の構成から、突発的な機械トラブル等に起因する突飛なデータが出現しない限り連続的に出現する。これにより、学習データ抽出処理部41は、特徴のある時系列データを追加学習データとして連続的に抽出することができ、このため、結果的に有効な学習データが連続的に出現する期間としての有効学習期間を抽出することが可能となる。

[0057] 続いて、リアルタイムの時刻での観測値が変動許容範囲を逸脱している旨を連続的に判定すると共に当該判定にかかる時系列データを連続的に抽出した学習データ抽出処理部41は、変動許容範囲を逸脱しない時系列データが出現した際に抽出した時系列データ（学習に必要である場合にはその近辺の時刻の時系列データも含めて）を、学習データ記憶部21Aに一括して記憶処理する（図2：S204）。

[0058] 次に、情報集計部61は、学習データ記憶部21Aに一括記憶された追加学習データが内包する予測属性情報を単位時刻に対応づけて学習集計データとして集計すると共に、これを学習処理部51に送信する（図2：S205）。

[0059] 続いて、学習処理部51は、情報集計部61から受け取った学習集計デー

タとの関係でデータ集中に対する各ノードの影響力を示す影響力データを算出すると共に（図2：S206）、この影響力データ及び当該算出に用いた学習集計データを学習処理情報保存部21Bに保存処理する（図2：S207）。

[0060] ここで、影響力データ及び学習集計データが学習処理情報保存部21B内に既に保存されていた場合には、学習処理部51がデータ更新処理機能51Aにて、リアルタイムで算出した影響力データ及び当該算出に用いた学習集計データにより学習処理情報保存部21B内の保存情報を更新する（図2：S207）。

[0061] また、学習処理情報保存部21B内に保存又は更新された学習集計データが存在する場合には、学習処理部51が再学習処理機能51Bにて、この保存された学習集計データとリアルタイムにて情報集計部61から取得する学習集計データとを結合処理すると共にこれを使用して影響力データを算出する（図2：S206）。上記同様に、データ更新処理機能51Aにて、この影響力データ及び当該算出に用いた学習集計データにより学習処理情報保存部21B内の保存情報を更新する（図2：S207）。

[0062] ここで、上述した学習データ抽出処理部41の抽出処理によって得られる追加学習データは、直近の過去数時刻の傾向と比較して特に特徴的な振る舞いを示す時刻に対応する時系列データに限定されることがあるため、データ変動の程度にかかわらず全ての知識を抽出する必要がある場合等には、当該抽出処理を実行しないか又はパラメータ（ U 、 α ）の調整等により制御するようにしてもよい。

[0063] また、学習データ抽出処理部41を、上記逸脱している旨の判定にかかる時刻が単一の時刻のみであった場合又は予め設定した連続時刻数を超えなかった場合には、当該判定にかかる時系列データを学習データ記憶部21Aに記憶処理しないような構成としてもよい。即ち、変動許容範囲を逸脱する旨の判定をある程度連続する時刻（ある程度の期間内）において行った場合にのみ、学習データ抽出処理部41が追加学習データとして記憶処理するよう

にしてもよい。このようにすれば、突発的な機械トラブル等に起因する有意性の認められないデータを排除することができるため、判定及びその後の予測の精度向上を図ることが可能となる。

[0064] (予測処理)

次に、図3に基づいてデータ集中の予測処理を説明する。

データ入力手段11が、予め設定された時間間隔（予測間隔）で発せられる予測要求に応じてデータ集中予測用の時系列データを取り込むと共に、これを情報集計部61に送信する（図3：S208）。続いて、情報集計部61は予測データ集計機能61Bにて、データ入力手段11から受信した時系列データに含まれる予測属性情報を単位時刻に対応づけて予測集計データとして集計すると共に、これを予測処理部71に送信する（図3：S209）。

[0065] 続いて、予測集計データを受け取った予測処理部71は、この予測集計データと学習処理部51が算出した影響力データとに基づいてデータ集中にかかる予測値を算出する。その際、必要に応じて予測処理部71は、算出した予測値をデータ記憶手段21に記憶処理する（図3：S210）。この予測処理部71は予測結果出力機能71Aにて予測結果を装置外部に送信する（図3：S211）。

[0066] 上記各ステップS201～S211（図2、図3）における各工程の実行内容をプログラム化すると共に、この一連の各制御プログラムをコンピュータによって実現するように構成してもよい。

[0067] (第1実施形態の効果)

本第1実施形態では、過去の観測値（予測属性情報）の変動傾向に基づいて特徴的な時系列データを抽出する構成を採用したため、これにより、観測値の多様な変動傾向に順応した必要十分な範囲内のデータを自動的に抽出できる。即ち、学習データ抽出処理部41は、図5に示すように、追加学習データを波形の幅や抑揚に順応した有効学習期間の全域に亘って自動的に抽出することができるため、これにより、過剰なデータ収集や必要なデータが不

足するといった事態を自動的に回避することが可能となる。

[0068] このようにすれば、例えば、自動車の故障、ツイッターやブログ上での炎上（ブログ等の記載に反応して多数の閲覧者がコメントを集中的に寄せる状態）といった特徴的な振る舞いを示す時刻にかかる有用な知識を、学習データ抽出処理部41にて効果的に抽出することができるため、これにより、当該抽出データを利用した精度のよいデータ集中の予測が可能となる。

[0069] [第2実施形態]

本発明の第2実施形態にかかるデータ集中予測装置を、図6乃至図9に基づいて説明する。

ここで、前述した第1実施形態と同一の構成部材については同一の符号を用いるものとする。この第2実施形態では、具体例としてツイッター（Twitter：登録商標）上に投稿されたテキストデータを分析対象とし、指定した話題（トピック）に関する未来の投稿件数（ツイート回数の総和）を予測する構成及び動作を説明する。

即ち、本第2実施形態では、予測属性情報（データ集中の予測にかかる属性のデータ）として投稿件数を採用し、リアルタイムの特定時刻から見たS時刻先の投稿件数を予測する場合を例示する。また、未来の投稿件数を予測したい話題（複数でも可）は、外部入力等により予め指定されているものとする。

[0070] （全体的構成）

本第2実施形態におけるデータ集中予測装置82は、図6に示すように、複数のノードND（1～n）としてのユーザ（発信者）から発信されるツイッターデータ（Twitterデータ）を対応する属性情報と共にネットワーク92を介して受信し時系列データとして取り込むデータ入力手段12と、この取り込んだ時系列データを記憶するデータ記憶手段21と、この記憶した時系列データのデータ構造を分析し後発するデータ集中を予測するデータ集中予測手段32と、を備えている。

[0071] ここでのツイッターデータとは、ツイッター上にツイート（投稿）された

テキストデータ及びこれと同時に入力される各情報（「ツイート時刻」・「ツイートしたノード」・「そのテキストデータが属する話題」にかかる各情報）を意味する。

[0072] データ入力手段１２は、経時的に学習データ抽出用のツイッターデータを入力する学習用データ入力部１２Ａと、各ツイッターデータに紐付けられたノードごとの属性情報を入力する属性情報入力部１２Ｂと、予め設定された時間間隔（予測間隔）で発せられる予測要求に応じてデータ集中予測用のツイッターデータ（予測データ）を入力する予測用データ入力部１２Ｃと、を備えている。

[0073] ここで、属性情報入力部１２Ｂは、ノードの属性情報として、ノードごとのTwitterクライアント、有効学習期間内のツイート回数、有効学習期間内でのコメント数・トラックバック数・リプライ数・リツイート数の各平均値、有効学習期間内でのフォロー数、及びフォロワー数の最大値等の情報を取得するように構成されている。

[0074] データ集中予測手段３２は、データ入力手段１２Ａが取り込むツイッターデータを予め設定された単位時刻（単位集計時間）ごとに区分して経時的に一時記憶処理すると共に、この一時記憶処理データを利用して各ツイッターデータの入力時刻の直前の時刻を基準とした過去の一定期間内のツイッターデータに基づいて予め設定される変動許容範囲を逸脱する当該ツイッターデータを後発するデータ集中の予測に必要な追加学習データとして連続的に抽出処理する学習データ抽出処理部４２を備えている。

[0075] データ記憶手段２１は、学習データ抽出処理部４２が連続的に抽出した追加学習データを記憶する学習データ記憶部２１Ａと、この記憶された追加学習データに基づいてデータ集中予測手段３２が行う各種算出処理の結果情報を記憶する学習処理情報保存部２１Ｂと、を含む構成から成る。学習データ抽出処理部４２は、連続的に抽出した追加学習データを学習データ記憶部２１Ａに一括して記憶処理する学習データ記憶処理機能４２Ａを有している。

[0076] また、データ集中予測手段３２は、学習データ記憶部２１Ａから取得した

追加学習データに基づいて過学習を抑止する正則化手法を取り入れた算出機能によりデータ集中に対するノードの影響力を示す影響力データを算出処理する学習処理部 5 2 と、属性情報入力部 1 2 B から取得したノードごとの属性情報を使用してデータ集計処理を実行する情報集計部 6 2 と、予め設定された時間間隔（予測間隔）で発せられる予測要求に応じて予測用データ入力部 1 2 C が経時的に取り込む予測データに基づいてデータ集中を予測する予測処理部 7 2 と、を備えている。

[0077] ここで、ノードの影響力とは、ノードがデータ集中の予測値に及ぼす影響力のことである。即ち、影響力データは、データの集中発信近辺の時刻において各ノードが当該集中発信にどの程度寄与したかを示すデータである。また、予め外部入力により設定された予測間隔にかかる情報はデータ記憶手段 2 1 に記憶されているものとする。本第 2 実施形態では、この予測間隔にかかる情報を予測処理部 7 2 がデータ記憶手段 2 1 から取得すると共に当該情報に沿ってデータ入力手段 1 2 に対し予測要求を行うように構成されている。

[0078] 学習処理部 5 2 は、学習データ記憶部 2 1 A から取得した追加学習データに含まれるテキストデータを話題ごとに分類し話題ごとの学習分類情報（その話題に属するツイッターデータのノード情報・時刻情報・テキストデータ）を作成する学習分類機能 5 2 A を有し、予測処理部 7 2 は、上記予測間隔で発せられる予測要求に応じて予測用データ入力部 1 2 C が経時的に取り込む予測データを話題ごとに分類し話題ごとの予測分類情報（その話題に属するツイッターデータのノード情報・時刻情報・テキストデータ）を作成する予測分類機能 7 2 A を有する。

[0079] 情報集計部 6 2 は、学習処理部 5 2 が作成する学習分類情報又は予測処理部 7 2 が作成する予測分類情報と、属性情報入力部 1 2 B から取得したノードごとの属性情報とに基づいて、全ノードのグループ化（グルーピング）を実行しグループ情報（各ノードが属するグループにかかる情報）を作成するグループ化機能 6 2 C を有する。

[0080] また、情報集計部 6 2 は、学習分類情報を利用してグループ化機能 6 2 C が作成したグループ情報を単位時刻に対応づけることにより学習集計データとしてのクロス集計データ（話題ごとのグループ・時刻情報に関するツイート数のクロス集計表）を作成する学習データ集計機能 6 2 A と、予測分類情報を利用してグループ化機能 6 2 C が作成したグループ情報を単位時刻に対応づけることにより予測集計データとしてのクロス集計データを作成する予測データ集計機能 6 2 B と、を備えている。ここで、情報集計部 6 2 が有する学習データ集計機能 6 2 A 及び予測データ集計機能 6 2 B を総称してグループ集計機能 6 3 と指称する。

[0081] 学習処理部 5 2 は、この学習データ集計機能 6 2 A が集計して成る学習集計データに基づいて、前述した影響力データの算出を行うように構成されている。

[0082] また、予測処理部 7 2 は前述したデータ集中の予測を、予測データ集計機能 6 2 B が集計して成る予測集計データと学習処理部 5 2 が算出する影響力データとに基づいて実行し、これにより、未来の投稿件数の予測値（後発するデータ集中にかかる予測値）を算出処理するように構成されている。この予測処理部 7 2 は、算出した未来の投稿件数の予測値を装置外部に出力する予測結果出力機能 7 1 A を有している。また、予測処理部 7 2 が、当該算出にかかる予測値をデータ記憶手段 2 1 に記憶処理するように構成してもよい。

[0083] 上記各ノードのグループ化は、例えば、「Twitterクライアントは何か？」「学習期間内でのツイート回数は1～100回、101～1000回、1001回以上のどれか？」「学習期間内でのフォロワー数の最大値は1～1000回、1001回以上のどれか？」といった属性の種類ごとに行われる。

また、グループ情報とは、各ノードの属性情報の共通性に基づいてグループを定めると共に各ノードを1以上のグループに帰属させた情報である。

[0084] 特に、ノードの数が不安定である場合には、グループの数を特定の数に限定することにより各ノードをこれらのグループに帰属させることができるた

め（実質的に見てノードの数を減少させ且つ安定させたことと同議であるため）、これにより、回帰分析等の統計的手法を用いた処理を迅速且つ高精度に行うことが可能となる。即ち、本第2実施形態では、このグループ化により、ノードの影響力の学習結果を安定させることができる。

[0085] さらに、必要に応じて、属性の種類ごとに行ったグループ化の結果の積集合に基づいて最終的なグループ情報を作成するようにしてもよい。また、ツイート回数がある一定値以上のノードについては、そのノード自体を単独のグループと定義してのグループ化を行うようにしてもよい。このようにすれば、ツイッター上で特に影響力のあるノードの情報を的確に把握することができる。

[0086] また、学習処理部52は、リアルタイムで算出した影響力データ及びその算出に用いた学習集計データに加工を施した情報（集計データ加工情報）を前処理データとして学習処理情報保存部21Bに保存処理する機能と共に、当該前処理データが既に保存されている場合には、リアルタイムで算出した影響力データ及びその算出にかかる集計データ加工情報により学習処理情報保存部21B内の前処理データとしての保存情報を更新するデータ更新処理機能51Aを備えている。更に、この学習処理部52は、同様に前処理データが既に保存されている場合にはリアルタイムでの影響力の算出に際して、この保存された過去の集計データ加工情報とリアルタイムでの集計データ加工情報とを結合処理すると共にこれを使用して影響力データを算出する再学習処理機能51Bと、算出したグループの影響力をもとに、各ノードが属するグループの影響力の総和を当該ノードの影響力としノードについての影響力データを算出する影響力処理機能52Bと、を備えている。

[0087] 以下、グループの数を一定数 G とし、数式を参照してデータ集中予測装置82の情報処理の内容を説明する（統計的手法により影響力データを算出し、後発するデータ集中を予測する手法等を開示する）。

[0088] 学習データ抽出処理部42は、学習用データ入力部12Aから取り込むと共に揮発性メモリに一時記憶させたツイッターデータに基づいて各時刻ごと

に定まる変動許容範囲を下記式 1 及び式 2 により算出し、リアルタイムの時刻における予測属性情報としての全ツイート回数（投稿件数）が当該変動許容範囲を逸脱しているか否かの判定を連続的に行うように構成されている。

[0089] [数1]

$$\hat{\mu}_{T'+1} + \alpha \hat{\sigma}_{T'+1} < y_{T'+1} \quad \cdots \quad (1)$$

[0090] [数2]

$$\hat{\mu}_{T'+1} = \frac{1}{U} \sum_{u=1}^U y_{T'-U+u} \quad , \quad \hat{\sigma}_{T'+1}^2 = \frac{1}{U} \sum_{u=1}^U (y_{T'-U+u} - \hat{\mu}_{T'+1})^2 \quad \cdots \quad (2)$$

[0091] 上記式 1 及び式 2 は、ある時刻 $T' + 1$ （リアルタイムの時刻）に観測されたツイッターデータの追加学習データ（後発するデータ集中を予測する際に必要な学習データ）としての適切性を評価する指標として用いられる。

ここで $y_{T'+1}$ は、時刻 $T' + 1$ の全ノードに関するツイート回数の総和（投稿件数）であり、

$$\hat{\mu}_{T'+1} \text{ と } \hat{\sigma}_{T'+1} \text{ (分散 } \hat{\sigma}_{T'+1}^2 \text{ の正の平方根)}$$

は、式 2 に示す通り、時刻 T' から遡って連続する合計 U の時刻（直近過去 U 時刻）のツイート回数に基づいて算出する平均値と標準偏差である。また、 U 及び逸脱に対する感応度 α は外部から入力されるパラメータである。

[0092] また、ここでは投稿件数の盛り上がりに着目するため、式 1 の左辺に示す変動許容範囲を大きい方に逸脱した投稿件数「 $y_{T'+1}$ 」にかかるツイッターデータを追加学習データとして学習データ抽出処理部 42 が抽出する構成を採用した。

[0093] 即ち、学習データ抽出処理部 42 は、リアルタイムの時刻での投稿件数が、過去の一定期間内（直近過去 U 時刻）の投稿件数の平均値を基準とした変動許容範囲内にあれば（式 1 を満たさなければ）追加学習データとして不適切である旨判定し、変動許容範囲から外れていれば（式 1 を満たせば）追加学習データとして適切である旨判定するように構成されている。

[0094] ここで、データ集中予測装置 82 においては、学習データ抽出処理部 42

が抽出処理した追加学習データとしてのツイッターデータに基づいて最終的に予測処理部 7 2 が未来の投稿件数の予測値を算出するように構成されている。したがって、追加学習データの抽出に用いる逸脱に対する感応度 α (式 1) は予測の精度を大きく左右する。そこで本第 2 実施形態においても、何パターンかの α の値に対して、予め記憶された過去のデータを利用した学習・予測処理を実行し、もっとも予測精度が良かった α の値を上記式 1 に用いることで予測精度の向上を図った。また、この α の値を予め決定する際の評価指標としては、平均 2 乗誤差を採用した。

[0095] また、学習データ抽出処理部 4 2 は、時刻 $T' + 1$ にかかるツイッターデータが追加学習データとして適切である旨判定すると、当該判定の時刻 $T' + 1$ にかかるツイッターデータを学習に必要な直近過去 S 時刻 ($T' + 1 - S, T' + 2 - S, \dots, T' + 1 - 1$) にかかるツイッターデータと共に抽出するように構成されている。

[0096] さらに、学習データ抽出処理部 4 2 は上記式 1 及び式 2 の構成から、追加学習データとして適切である旨の判定を突飛なデータが出現しない限り連続的に行い、直近の過去数時刻の投稿件数と比較して大きな増加を示す時刻にかかるツイッターデータ (直近過去 S 時刻にかかるツイッターデータも含む) を追加学習データとして学習データ記憶部 2 1 A にまとめて一回で記憶処理するように構成されている。

[0097] 学習データ集計機能 6 2 A は、学習処理部 5 2 から取得した学習分類情報とこれに対応するグループ化機能 6 2 C から入力したグループ情報に基づいて、各時刻・各グループにおけるツイート回数を集計し、下記行列式 3 で表わされる学習集計データ (クロス集計データ) を作成するように構成されている。

[0098] [数 3]

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1G} \\ x_{21} & x_{22} & \cdots & x_{2G} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{TG} \end{pmatrix} \quad \cdots \quad (3)$$

ここで、

$$t = 1, 2, \dots, T$$

$$g = 1, 2, \dots, G$$

[0099] この式 3 に示す行列において、各行は各時刻を表し、各列は各グループを表す。また、各要素 x_{tg} の値は時刻ごとの投稿件数を表し、各グループのツイート回数の総和が時刻ごと整理されている。

[0100] 即ち、学習データ集計機能 6 2 A は、予め指定された単位時刻とグループ情報に従い、学習分類情報に対して「どの時刻においてどのグループが何回ツイートしたか」についての集計作業を行うことにより学習集計データを作成するように構成されている。

[0101] 本第 2 実施形態では、学習データ抽出処理部 4 2 が新たな追加学習データを抽出処理した際に学習処理部 5 2 が、ノードごとの影響力を示す影響力データを算出すると共に、この影響力データと当該算出に用いた学習集計データに加工を施した情報である各時刻・各グループの集計結果 X_s, y_s ($s = 1, 2, \dots, S$) (集計データ加工情報) を学習処理情報保存部 2 1 B に保存処理しておくように構成されている。

ここで、 X_s は、式 3 の形式で示した X の第 1 行目から第 $T - s$ 行目までを抜き出した行列であり、 $y_s = (y_{s+1}, \dots, y_T)'$ は、各時刻におけるツイート回数の全ノードに関する総和 (投稿件数) である。

[0102] 本第 2 実施形態では、グループの影響力を示す影響力データが下記式 4 に示す行列で与えられると仮定する。また、ノードごとの影響力は、グループの影響力に基づいて算出される。

[0103] [数 4]

$$\beta = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1G} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2G} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{S1} & \beta_{S2} & \cdots & \beta_{SG} \end{pmatrix} \cdots \quad (4)$$

ここで、

$$s = 1, 2, \dots, S$$

$$g = 1, 2, \dots, G$$

[0104] [数5]

$$f(y_s, X_s, \beta_s) + \lambda P(\beta_s), \quad (s = 1, 2, \dots, S) \quad \cdots \quad (5)$$

[0105] 上記式4に示す行列において、各行は単位時刻（単位集計時間）のもとで何時刻先の未来であるか（ある基準時刻からみたS時刻先の未来までの時刻）を表わし、各列は各グループを表す。また、行列の各要素 β_{sg} の値は各グループの影響力を表わす。

[0106] ここでは学習処理部52が、上記式5を最小化する β をグループの影響力を示す影響力データとして算出するように構成されている。 λ は正則化パラメータと指称される学習結果の安定性を調整するパラメータである。

[0107] このとき、ノードの影響力は、そのノードが属するグループの影響力の総和として定義することができる。

即ち、学習処理部52は影響力処理機能52Bにより、この算出したグループの影響力 β に基づいて、各ノードが属するグループの影響力の総和を当該ノードの影響力として算出するように構成されている。ここで、学習処理部52が各グループの影響力に基づいてノードごとの影響力を算出する手法を図9に基づいて説明する。

[0108] 図9は、ノードND(1)～ND(n)のグループ化としてグループGP(1)～GP(3)の3つのグループに分けるという例を示す。ノードND(1)、ND(2)は、グループGP(1)にのみ帰属している。ノードND(4)、ND(7)は、それぞれグループGP(2)、GP(3)にのみ帰属している。また、ノードND(3)は、グループGP(1)とGP(2)の双方に帰属し、ノードND(6)は、グループGP(2)とGP(3)の双方に帰属し、ノードND(5)は、グループGP(1)からGP(3)のすべてに帰属している。

[0109] この場合、ノードND(1)、ND(2)の影響力は、グループGP(1)の影響力と等しくなり、ノードND(4)、ND(7)の影響力は、それぞれグループGP(2)、GP(3)の影響力と等しくなる。

[0110] 一方で、グループGP(1)とGP(2)の双方に帰属しているノードND(3)の影響力

は、グループGP(1)とGP(2)のそれぞれの影響力の合計として算出される。同様に、ノードND(6)の影響力は、グループGP(2)とGP(3)のそれぞれの影響力の合計として算出される。また、ノードND(5)の影響力は、グループGP(1)からGP(3)のすべての影響力を加算した値として算出される。即ち、複数のグループに属するノードの影響力が大きくなるように構成されている。

[0111] (正則化機能)

また、過学習を抑止する正則化手法を取り入れた算出機能を有する学習処理部52は、上記式5の要素として、下記式6及びL1正則化を示す式7又はL2正則化を示す式8を採用しての処理を実行するように構成されている。P。(x, α)は、平均αのポアソン分布のxにおける密度関数の値を示す。

[0112] [数6]

$$f(y_s, X_s, \beta_s) = - \sum_{t=s+1}^T \log \left(Po \left(y_t, \exp \left(\sum_{g=1}^G \beta_{sg} x_{(t-s)g} \right) \right) \right) \quad \cdots (6)$$

[0113] [数7]

$$P(\beta_s) = \sum_{g=1}^G |\beta_{sg}| \quad \cdots (7)$$

[0114] [数8]

$$P(\beta_s) = \sum_{g=1}^G \beta_{sg}^2 \quad \cdots (8)$$

[0115] (結合処理)

ここで、学習処理部52による保存処理又は更新処理が行われた後の学習処理情報保存部21Bは、上述した通り、前処理データとして過去の集計データ加工情報である X_s 及び y_s ($s = 1, 2, \dots, S$) が保存された状態となるため、かかる場合において学習データ集計機能62Aからリアルタイムでの学習集計データ (上記式3の形式でクロス集計された情報) を受け取った学習処理部52は、当該前処理データを利用しての影響力データの算出処理を再学習処理機能51Bによって行うように構成されている。

[0116] 本第2実施形態では、学習データ抽出処理部42により新たな有効学習期間が抽出された際に学習処理部52は、当該有効学習期間内の追加学習データに基づく集計データ加工情報である

$$\tilde{X}_s \text{ 及び } \tilde{y}_s$$

を算出処理すると共に、学習処理情報保存部21Bから過去の集計データ加工情報である X_s 及び y_s を取得し、このリアルタイムで算出処理した

$$\tilde{X}_s \text{ 及び } \tilde{y}_s$$

と前処理データとしての X_s 及び y_s とを下記式9に示す形式によって結合処理するように構成されている。

[0117] [数9]

$$(X'_s, \tilde{X}'_s), (y'_s, \tilde{y}'_s) \cdots \quad (9)$$

[0118] また、この結合処理した式9に示す内容を上記式5に代入することで学習処理部52は、上記同様にして影響力データを算出すると共にこれを学習処理情報保存部21Aに記憶処理するように構成されている。このようにすれば、既存の学習データに関する集計のための計算（ X_s 及び y_s を得るための前処理の計算）を省略することできるため、結果的にノードの影響力の新たな学習を効率よく実現できる。

[0119] このとき、新たな学習データが追加されたことにより、ノードごとのグループ化の結果が変化する場合は、既存の学習データに関する学習集計データを算出し直すようにしてもよい。

[0120] (投稿件数の予測)

ここで、投稿件数の予測にかかる集計情報部62及び予測処理部72の構成を説明する。

本第2実施形態では、学習処理部52が算出する影響力データと別途取得した予測データとに基づく予測処理を予測処理部72が実行するように構成されている。

[0121] 例えば、予測データに関して、追加学習データと同様にグループ化及び集計処理を行った結果を $z = (z_1, \dots, z_G)$ と表わした場合、予測処理部 72 は s 時刻先の未来における投稿件数を下記式 10 によって予測するように構成されている。

[0122] [数10]

$$\sum_{g=1}^G \beta_{sg} z_g \quad \dots \quad (10)$$

[0123] これは、ある時刻から見て未来の投稿件数を予測する際に、その時刻における投稿件数のみを利用することを前提としたものである。

[0124] 次に、未来の投稿件数を予測するに際し、ある単一の時刻のみではなく、当該時刻を含む直近数時刻の投稿件数を利用するように拡張した場合について説明する。

ここでは、ある時刻から見ての未来の投稿件数を予測する際、当該ある時刻における投稿件数を含む直近の過去数時刻の投稿件数を利用する。例えば、ある時刻を含む直近過去 A 時刻の投稿件数を利用する場合には、予測データ集計機能 62B にて作成される予測集計データは下記式 11 の形式で表わされる。

[0125] [数11]

$$Z = \begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1G} \\ Z_{21} & Z_{22} & \dots & Z_{2G} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{A1} & Z_{A2} & \dots & Z_{AG} \end{pmatrix} \quad \dots \quad (11)$$

[0126] 予測処理部 72 は、この予測集計データを利用して、上記式 10 の形式により未来の投稿件数の予測値を算出するように構成されている。リアルタイムの時刻での予測を行う場合、予測データの取得及び予測処理は予め指定された時間間隔（予測間隔）で行う。

[0127] 時々刻々と変化するノードの影響力や、時間の経過に伴って影響力を持ったノードが新たに出現する可能性を考慮して、本第 2 実施形態では、追加学

習データの取得および前処理データの更新と共にノードの影響力の再学習を定期的に行う構成を採用した。このようにすれば、前処理により更新された適切なデータを再学習に利用することができるため、未来の投稿件数データにかかる出現予測を迅速に且つ精度よく実現することができる。

[0128] (動作説明)

次に、図6に示したデータ集中予測装置の動作制御の内容を図7及び図8に示すフローチャートに基づいて説明する。ここでは、煩雑さを避けるため、単一の話題に関する学習・予測について説明する。話題が複数指定されている場合は、各話題に対して以下で説明する手法と同じ手法で学習・予測を行う。

[0129] (学習処理)

まず、図7に基づいて時系列データの学習処理を説明する。

学習用データ入力部12Aに、複数のノードND(1~n)としてのユーザ(発信者)によりツイッター上にツイートされたテキストデータがネットワーク92を介して入力される。この際、各テキストデータに関する「ツイートされた時刻」「ツイートしたノード」及び「そのテキストデータが属する話題」にかかる各情報も同時に入力される。上述した通り、これらの情報から成るツイッターデータを学習用データ入力部12Aは、学習データ抽出処理部42に送信する(図7:S701)。

[0130] 続いて、学習用データ入力部12Aからツイッターデータを受け取った学習データ抽出処理部42は、上記式1及び式2に基づいて各単位時刻(各単位集計時間)におけるツイッターデータが追加学習データとして適切であるか否かを連続的に判定する(図7:S702)。

[0131] 具体的には、リアルタイムの時刻に学習用データ入力部12Aから受け取ったツイッターデータの投稿件数(予測属性情報)に基づいて、学習データ抽出処理部42が上記式1及び式2に基づく判定を実行し(図7:S702)、追加学習データとして適切である旨の判定にかかるツイッターデータを直近過去S時刻のツイッターデータと共に連続的に抽出する(図7:S70

3)。

[0132] 即ち、式1を具備する場合において学習データ抽出処理部42は、リアルタイムの時刻にかかるツイッターデータが追加学習データとして適切である旨判定すると共に（図7：S702／はい）当該判定にかかるツイッターデータ（直近過去S時刻にかかるツイッターデータも含む）を抽出し、後続するツイッターデータの抽出処理に移行する（図7：S703）。

[0133] 一方で、式1を具備しなければ、学習データ抽出処理部42は、そのツイッターデータが追加学習データとしては不適切である旨判定し、当該判定にかかるツイッターデータを抽出せずに後続するツイッターデータの処理に移行する（図7：S702／いいえ）。

[0134] 続いて、上記により追加学習データとして適切である旨を連続的に判定した学習データ抽出処理部42は、当該連続判定にかかる有効学習期間内のツイッターデータ（直近過去S時刻かかるデータも含む）を学習データ記憶部21Aにまとめて一回で記憶処理する。過去の学習データが学習データ記憶部21A内に存在するときは追加的に記憶させる（図7：S704）。

[0135] 学習データ抽出処理部42が有効学習期間内の新たなツイッターデータをデータ記憶部21Aに一括して記憶処理すると（図7：S704）、学習処理部52は、追加学習データとしての当該新たなツイッターデータをデータ記憶部21Aから取得する（図7：S705）。

[0136] 次いで、学習処理部52は、この新たなツイッターデータをノード情報・時刻情報・テキストデータの3つの情報に基づいて話題ごとに分類する。この分類により得た学習分類情報（単一の話題に属するツイッターデータの情報）を学習処理部52は、話題ごとに情報集計部62へと送信する（図7：S705）。

[0137] 続いて、学習用データ入力部12Aに入力される各テキストデータに紐付けられたノードごとの属性情報を受信した属性情報入力部12Bが、当該属性情報をグループ化機能62Cに送信する（図7：S706）。

[0138] グループ化機能62Cは、属性情報入力部12Bから取得したノードごと

の属性情報に基づいて有効学習期間内に受信したツイッターデータにかかる全ノードのグループ化（グルーピング）を実行し、これにより作成したグループ情報を学習データ集計機能62Aに送信する（図7：S707）。

[0139] 続いて、情報集計機能62Aは、学習処理部52から取得した学習分類情報とグループ化機能62Cから受信したグループ情報に基づいて、上記式3の形式による学習集計データ（クロス集計データ）を作成すると共に、これを学習処理部52に送信する（図7：S708）。

[0140] 続いて、学習データ集計機能62Aから学習集計データを受け取った学習処理部52は、これまでに作成され予め保存されている過去の集計データ加工情報を学習処理情報保存部21Bから取得し、これらの情報を再学習処理機能51Bにて上記式9の形式でまとめた後、上記式4の形式でグループの影響力を算出し、その値をもとに影響力処理機能52Bにて各ノードが属するグループの影響力の総和を当該ノードの影響力とし影響力データを算出する（図7：S709）。

[0141] 一方で、影響力データの算出に際して、学習処理情報保存部21Bに過去の集計データ加工情報が保存されていなかった場合には、リアルタイムで学習データ集計機能62Aから受け取った学習集計データに基づき上記式4の形式でグループの影響力を算出し、その値をもとに影響力処理機能52Bにてノードの影響力を示す影響力データを算出する（図7：S709）。

[0142] グループの影響力の算出に際して、学習処理部52は、上記式6及び上記L1正則化を示す式7又はL2正則化を示す式8を利用して上記式5の値を最小にするグループの影響力 β を、上記式4に示す形式で算出する。この学習処理部52の正則化手法を取り入れた算出機能によれば過学習を抑止することができるため、これにより、学習結果の安定性を向上させることが可能となる。（図7：S709）。

[0143] 続いて、学習処理部52はデータ更新処理機能51Aにて、リアルタイムで算出した影響力データ及びその算出にかかる集計データ加工情報により学習処理情報保存部21B内の情報を更新する。また、学習処理情報保存部2

1 B内に保存情報が存在しなかった場合には、学習処理部5 2はリアルタイムで算出した影響力データ及びその算出にかかる集計データ加工情報を学習処理情報保存部2 1 Bに保存処理する（図7：S 7 1 0）。

[0144] （投稿件数の予測）

次に、図8に基づいて、予測データが観測された時刻から見た未来の投稿件数を予測する一連の動作内容（将来の投稿件数にかかるデータ集中の予測処理）を説明する。

[0145] 予測用データ入力部1 2 Cが、予め設定された時間間隔（予測間隔）で発せられる予測要求に応じてデータ集中予測用のツイッターデータを入力する。この際、各テキストデータに関しての「ツイートされた時刻」「ツイートしたノード」及び「そのテキストデータが属する話題」にかかる各情報も併せて入力する。予測用データ入力部1 2 Cは、これらの情報から成る予測データとしてのツイッターデータを予測処理部7 2に送信する（図8：S 7 1 1）。

[0146] 続いて、予測処理部7 2は、取得した各ツイッターデータをノード情報・時刻情報・テキストデータの3つの情報に基づいて話題ごとに分類する。この分類により得た予測分類情報（単一の話題に属するツイッターデータの情報）を予測処理部7 2は、話題ごとに情報集計部6 2に送信する（図8：S 7 1 2）。

[0147] 続いて、予測用データ入力部1 2 Cが受信する各テキストデータに紐付けられたノードごとの属性情報を、属性情報入力部1 2 Bが装置外部から取り込むと共にこの属性情報を情報集計部6 2に送信する（図8：S 7 1 3）。

[0148] 情報集計部6 2はグループ化機能6 2 Cにて、属性情報入力部1 2 Bから取得したノードの属性情報に基づいてノードのグループ化を実行し、作成したグループ情報を情報集計部6 2に送信する（図8：S 7 1 4）。

[0149] 続いて、情報集計部6 2は、予測処理部7 2から取得した予測分類情報とグループ作成部6 2 Cから受信したグループ情報に基づいて話題ごとの予測集計データ（クロス集計データ）を上記式1 1の形式で作成すると共に、こ

れを予測処理部 7 2 に送信する（図 8 : S 7 1 5）

[0150] 続いて、学習処理部 5 2 が、予め学習処理情報保存部 2 1 B に保存処理された式 4 の形式から成る影響力データを取得すると共にこれを予測処理部 7 2 に送信すると、予測処理部 7 2 は、この影響力データ（式 4 の形式）と情報集計部 6 2 から受け取った予測集計データ（式 1 1 の形式）とを利用して、予測データが観測された時刻から見た未来の投稿件数を予測する。即ち、予測処理部 7 2 は、上記式 1 0 の形式で未来の投稿件数の予測値を算出する。その際、必要に応じて予測処理部 7 2 は、当該算出にかかる予測値をデータ記憶手段 2 1 に記憶処理する（図 8 : S 7 1 6）。

[0151] 次いで、予測処理部 7 2 は、算出した未来の投稿件数の予測値を予測結果出力機能 7 1 A により装置外部へと出力する（図 8 : S 7 1 7）。このようにすれば、かかる予測値を取得したツイッタークライアントやノード等が将来のデータ変動にかかる予測値を把握すると共に適宜何らかの手立てを講じ得るため、これにより、ツイッターでの炎上等の問題を未然に防止することが可能となる。

[0152] また、上記各ステップ S 7 0 1 ~ S 7 1 7（図 7、図 8）における各工程の実行内容をプログラム化すると共に、この一連の各制御プログラムをコンピュータによって実現するように構成してもよい。

[0153] （第 2 実施形態の効果）

[0154] 本第 2 実施形態においても学習データ抽出処理部 4 2 は、上記式 1 及び式 2 に基づく手法により、多様なデータの変動傾向に順応した有効学習期間と共に当該期間内のツイッターデータを抽出することができるため、これにより、データの過剰収集やデータ不足といった問題を自動的に防止することが可能となる。

[0155] また、データ集中予測装置 8 2 においては、学習データ抽出処理部 4 2 が抽出したデータ集中の予測に有意な追加学習データに対して集計処理等を施すと共に、当該処理後のデータに基づいて予測処理部 7 2 が将来の投稿件数を予測する。即ち、精度よく抽出したデータ集中予測用の追加学習データを

元のデータとして用いることにより、信頼性の高い予測値を得ることができ
るため、これにより、ツイッターでの炎上の原因となるノードの影響力等を
精度よく把握でき且つ風評被害等を未然に防止することが可能となる。

[0156] さらに、グループ化機能62Cによるノード（発信者）のグループ化によ
れば、不安定なノード（発信者）の数を予め設定したグループ数（定数）に
限定することができ、このため、クロス集計データや影響力データ等の算出
を円滑に行うと共に当該算出結果の安定化を図ることが可能となる。また、
学習処理部52に採用した正則化手法を取り入れた算出機能によれば、過学
習を抑止することができるため、これにより、精確かつ安定感のあるノード
の影響力を繰り返し得ることができる。

[0157] さらに、本第2実施形態にかかるデータ集中予測装置82によれば、有効
学習期間の抽出や再学習等を自動的に効率よく行うことができるため、これ
により、予測処理の迅速化や予測精度の向上を図ることが可能となる。従っ
て、かかる自動化により、特に実応用の場面においての人的コストや人為的
ミスの削減が実現可能となる。

[0158] [構成等にかかる応用例]

[0159] 本第2実施形態では、算出処理の迅速化やデータの安定化等の観点から、
データ集中予測装置82の有するグループ化機能62C及び学習処理部52
の有する正則化手法を取り入れた算出機能の両方を導入した場合を説明した
が、何れか一方のみを導入するように構成してもよい。

[0160] また、本第2実施形態にかかるデータ集中予測装置82では、学習処理部
52が影響力データを学習処理情報保存部21Bから取得すると共にこれを
予測処理部72に送信する構成を採用した（図8：S716）。しかし、こ
の影響力データは、予測処理部72が直接学習処理情報保存部21Bから取
得するように構成してもよい。

[0161] さらに、予め外部入力により設定された予測間隔にかかる情報は、データ
記憶手段21に併設されたメモリ内に記憶されるような構成としてもよい。
即ち、データ入力手段12が、予測間隔にかかる情報を取得しこれに沿って

予測データの入力を行うようにしてもよい。

[0162] また、本第2実施形態では、ツイッター上での炎上等の特徴的なデータ変動を把握することを主眼としていることから、投稿件数の上昇傾向を有効に把握するために上記式1を採用し、これに基づいて変動許容範囲を算出すると共に追加学習データを抽出した（図7：S702、S703）。しかし、データの上昇傾向だけではなく減少傾向についても精度よく把握したい場合には、上記式1と共に下記式12を採用するようにしてもよい。

[0163] [数12]

$$\hat{\mu}_{T'+1} - \alpha \hat{\sigma}_{T'+1} > y_{T'+1} \quad \cdots \quad (12)$$

[0164] もっとも、データの減少傾向のみを有効に把握したい場合には、上記式1に代えて式12を採用するようにしてもよい。このようにすれば、各種時系列データの特徴や把握したい状況に応じての柔軟なデータ抽出が可能となるため、これにより、様々な場面でのデータ予測が実現可能となる。

[0165] また、本第2実施形態では、時系列データとしてツイッターデータ及びそれに紐付けられた属性情報を採用しての構成及び動作を説明した。しかし、本発明にかかるデータ集中予測装置82によれば、多様な環境下において現れる様々な時系列データについて、上述した各工程内容（図7及び図8：S701～S717）と同様の動作制御を実現できる。即ち、ツイッターやブログ等のWeb上のソーシャルメディアへの投稿記事数のみならず、地震の波形や津波の潮位変動といった自然現象にかかるデータ、自動車や工場ラインに設置されたセンサから得られる各部品の状態にかかるデータ、日々の生活における電力使用量といった人間活動にかかるデータ等の未来の出現予測を精度よく行うことができる。

[0166] なお、上述した実施形態は、データ集中予測装置、データ集中予測方法、及びそのプログラムにおける好適な具体例であり、技術的に好ましい種々の限定を付している場合もある。しかし、本発明の技術範囲は、特に本発明を限定する記載がない限り、これらの態様に限定されるものではない。

[0167] 以下は、上述した実施形態についての新規な技術的内容の要点をまとめた

ものであるが、本発明は必ずしもこれに限定されるものではない。

[0168] (付記 1)

複数のノードから発信されるデータを対応する属性情報と共に受信し時系列データとして取り込むデータ入力手段と、この取り込んだ時系列データを学習データとして記憶するデータ記憶手段と、前記記憶した時系列データのデータ構造を分析し後発するデータ集中を予測するデータ集中予測手段と、を備えたデータ集中予測装置であって、

前記データ集中予測手段は、前記データ入力手段が取り込む時系列データを予め設定された単位時刻ごとに経時的に一時記憶処理すると共に、この一時記憶処理データを利用して各時系列データの入力時刻の直前の時刻を基準とした過去の一定期間内の時系列データに基づいて予め設定される変動許容範囲を逸脱する当該時系列データを前記後発するデータ集中の予測に必要な追加学習データとして連続的に抽出処理する学習データ抽出処理部を備え、

この学習データ抽出処理部が、前記連続的に抽出した追加学習データを前記データ記憶手段に一括して記憶処理する学習データ記憶処理機能を備えたことを特徴とするデータ集中予測装置。

[0169] (付記 2)

前記付記 1 に記載のデータ集中予測装置において、

前記追加学習データの抽出に際して設定される変動許容範囲を、前記過去の一定期間内の時系列データが内包する前記データ集中の予測にかかる属性のデータの平均値と分散に基づいて前記学習データ抽出処理部が算出し設定することを特徴としたデータ集中予測装置。

[0170] (付記 3)

前記付記 1 又は 2 に記載のデータ集中予測装置において、

前記データ集中予測手段が、

前記データ記憶手段に一括記憶された追加学習データが内包する前記データ集中の予測にかかる属性のデータを前記単位時刻に対応づけて学習集計データとして集計する学習データ集計機能を有する情報集計部と、

前記データ集中にかかる予測値に対する各ノードの影響力を示す影響力データを前記学習集計データとの関係で算出すると共に、この影響力データ及び当該算出に用いた学習集計データを前記データ記憶手段に保存処理する学習処理部と、を備えたことを特徴とするデータ集中予測装置。

[0171] (付記4)

前記付記3に記載のデータ集中予測装置において、

前記情報集計部は、更に、予め設定された時間間隔で発せられる予測要求に応じて前記データ入力手段が取り込んだ時系列データに含まれる前記予測にかかる属性のデータを前記単位時刻に対応づけて予測集計データとして集計する予測データ集計機能を備え、

前記データ集中予測手段が、更に、前記予測集計データと前記影響力データとに基づいて前記予測値を算出処理する予測処理部を備えたことを特徴とするデータ集中予測装置。

[0172] (付記5)

前記付記3又は4に記載のデータ集中予測装置において、

前記学習処理部は、更に、リアルタイムでの前記影響力データの算出に際して前記データ記憶手段に予め影響力データ及び学習集計データが保存されていた場合には、リアルタイムにて算出した影響力データ及び当該算出に用いた学習集計データにより前記データ記憶手段内の保存情報を更新するデータ更新処理機能を備えたことを特徴とするデータ集中予測装置。

[0173] (付記6)

前記付記3又は4に記載のデータ集中予測装置において、

前記学習処理部は、更に、リアルタイムでの前記影響力データの算出に際して前記データ記憶手段に予め学習集計データが保存されていた場合には、その保存された学習集計データとリアルタイムにて前記情報集計部から取得する学習集計データとを結合処理すると共にこれを使用して前記影響力データを算出する再学習処理機能を備えたことを特徴するデータ集中予測装置。

[0174] (付記7)

前記付記 5 に記載のデータ集中予測装置において、

前記学習処理部は、更に、リアルタイムでの前記影響力データの算出に際して前記データ記憶手段に予め学習集計データが保存されていた場合には、その保存された学習集計データとリアルタイムにて前記情報集計部から取得する学習集計データとを結合処理すると共にこれを使用して前記影響力データを算出する再学習処理機能を備えたことを特徴するデータ集中予測装置。

[0175] (付記 8)

前記付記 7 に記載のデータ集中予測装置において、

前記学習処理部は、前記影響力データの算出に用いた学習集計データを前記データ記憶手段に保存処理する際又は前記データ更新処理機能にて前記データ記憶手段の保存情報を更新する際に、当該学習集計データに所定の加工を施した集計データ加工情報を用いて前記保存処理及び更新を実行し、

前記再学習処理機能は、前記集計データ加工情報とリアルタイムにて取得した学習集計データに前記所定の加工を施したものとを結合処理することを特徴としたデータ集中予測装置。

[0176] (付記 9)

前記付記 4 乃至 8 の何れか一つに記載のデータ集中予測装置において、

前記情報集計部は、

前記各ノードの属性データの共通性に基づいてグループを定めると共に各ノードを 1 以上のグループに帰属させてグループ情報を作成するグループ化機能と、

このグループ情報を前記予測にかかる属性のデータに代えて、前記時刻に対応づけることにより前記学習集計データ又は前記予測集計データを作成するグループ集計機能と、を備えたことを特徴とするデータ集中予測装置。

[0177] (付記 10)

前記付記 9 に記載のデータ集中予測装置において、

前記学習処理部は、前記各ノードにかかる影響力データを算出するに際して、前記データ集中に対する各グループの影響力を前記学習集計データとの

関係で算出すると共に、前記各ノードについて当該ノードの属するグループの影響力の加算値を当該ノードの影響力とする影響力処理機能を備えたことを特徴とするデータ集中予測装置。

[0178] (付記 1 1)

前記付記 3 乃至 1 0 の何れか一つに記載のデータ集中予測装置において、
前記学習処理部は、更に、前記データ記憶手段に一括記憶された追加学習データをデータ内容ごとに分類しこれにより作成した学習分類情報を前記情報集計部に送信する学習分類機能を備え、

前記情報集計部は、この学習分類情報を利用して前記学習集計データを作成することを特徴としたデータ集中予測装置。

[0179] (付記 1 2)

前記付記 4 乃至 1 1 の何れか一つに記載のデータ集中予測装置において、
前記予測処理部は、更に、前記予測要求に応じて前記データ入力手段が取り込んだ時系列データをデータ内容ごとに分類しこれにより作成した予測分類情報を前記情報集計部に送信する予測分類機能を備え、

前記情報集計部は、この予測分類情報を利用して前記予測集計データを作成することを特徴としたデータ集中予測装置。

[0180] (付記 1 3)

前記付記 4 乃至 1 2 の何れか一つに記載のデータ集中予測装置において、
前記予測処理部は、更に、将来のデータ変動の傾向として前記算出した予測値を装置外部に出力する予測結果出力機能を備えたことを特徴とするデータ集中予測装置。

[0181] (付記 1 4)

前記付記 3 乃至 1 3 の何れか一つに記載のデータ集中予測装置において、
前記学習処理部は、過学習を抑止する正則化手法を取り入れた算出機能により前記影響力データを算出することを特徴としたデータ集中予測装置。

[0182] (付記 1 5)

前記付記 1 4 に記載のデータ集中予測装置において、

前記学習処理部は、前記正則化手法としてL 1 正則化を実装したことを特徴とするデータ集中予測装置。

[0183] (付記 1 6)

前記付記 1 4 に記載のデータ集中予測装置において、

前記学習処理部は、前記正則化手法としてL 2 正則化を実装したことを特徴とするデータ集中予測装置。

[0184] (付記 1 7)

複数のノードから発信されるデータを対応する属性情報と共に受信し時系列データとして取り込むデータ入力手段と、この取り込んだ時系列データを学習データとして記憶するデータ記憶手段と、前記記憶した時系列データのデータ構造を分析し後発するデータ集中を予測するデータ集中予測手段と、を備えたデータ集中予測装置にあって、

前記データ集中予測手段は、前記データ集中の予測用の時系列データを抽出し処理する学習データ抽出処理部を備えており、

前記データ入力手段にて取り込まれた時系列データを予め設定された単位時刻ごとに経時的に一時記憶処理し、

この一時記憶処理データを利用して各時系列データが各々の入力時刻の直前の時刻を基準とした過去の一定期間内の時系列データに基づいて設定される変動許容範囲を逸脱するものであるか否かを判定し、

前記変動許容範囲を逸脱する旨判定した際に、当該判定にかかる時系列データを前記後発するデータ集中の予測に必要な追加学習データとして連続的に抽出し、

この連続的に抽出した追加学習データを前記データ記憶手段に一括して記憶処理すると共に、

これら一連の各工程内容を前記学習データ抽出処理部が順次実行し、

前記データ記憶手段に一括記憶された追加学習データ及び既存の学習データにより特定される前記時系列データのデータ構造に基づいて、前記データ集中予測手段の予測処理部が前記データ集中を予測することを特徴としたデ

ータ集中予測方法。

[0185] (付記 18)

前記付記 17 に記載のデータ集中予測方法において、

前記学習データ抽出処理部が前記連続的に抽出した追加学習データを一括して記憶処理した後、前記予測処理部による予測に先立って、

この一括記憶された追加学習データが内包する前記データ集中の予測にかかる属性のデータを前記単位時刻に対応づけて学習集計データとして集計し、

この学習集計データとの関係で前記データ集中にかかる予測値に対する各ノードの影響力を示す影響力データを算出し、

この影響力データ及び当該算出に用いた学習集計データにより前記データ記憶手段内の保存情報を更新し、

これら一連の各工程内容を前記データ集中予測手段が順次実行することを特徴としたデータ集中予測方法。

[0186] (付記 19)

複数のノードから発信されるデータを対応する属性情報と共に受信し時系列データとして取り込むデータ入力手段と、この取り込んだ時系列データを学習データとして記憶するデータ記憶手段と、前記記憶した時系列データのデータ構造を分析し後発するデータ集中を予測するデータ集中予測手段と、を備えたデータ集中予測装置にあって、

前記データ入力手段にて取り込まれた時系列データを予め設定された単位時刻ごとに経時的に一時記憶処理するデータ一時記憶処理機能、

この一時記憶処理データを利用して各時系列データが各々の入力時刻の直前の時刻を基準とした過去の一定期間内の時系列データに基づいて設定される変動許容範囲を逸脱するものであるか否かを判定する変動許容判定機能、

前記変動許容範囲を逸脱する旨判定した際に、当該判定にかかる時系列データを前記後発するデータ集中の予測に必要な追加学習データとして連続的に抽出する学習データ抽出機能、

この連続的に抽出した追加学習データを前記データ記憶手段に一括して記憶処理する学習データ記憶処理機能、

および、この一括記憶された追加学習データ及び既存の学習データにより特定される前記時系列データのデータ構造に基づいて前記データ集中を予測する予測処理機能を設け、

これら各情報処理機能を、前記データ集中予測手段が備えているコンピュータに実現させるようにしたことを特徴とするデータ集中予測プログラム。

[0187] (付記 20)

前記付記 19 に記載のデータ集中予測プログラムにおいて、

前記データ記憶手段に一括記憶された追加学習データが内包する前記データ集中の予測にかかる属性のデータを前記単位時刻に対応づけて学習集計データとして集計する学習データ集計機能、

この学習集計データとの関係で前記データ集中にかかる予測値に対する各ノードの影響力を示す影響力データを算出する影響力データ算出機能、

この影響力データ及び当該算出に用いた学習集計データにより前記データ記憶手段の保存情報を更新するデータ更新処理機能、

この更新された前記既存の学習データとしての保存情報を利用してリアルタイムで受信した時系列データに基づく前記予測値を算出すると共にこれを前記データ記憶手段に格納する予測値算出格納機能を設け、

これら各情報処理機能を、前記コンピュータに実現させるようにしたことを特徴とするデータ集中予測プログラム。

[0188] この出願は 2012 年 9 月 12 日に提出された日本出願特願 2012-200440 を基礎とする優先権を主張し、その開示の全てをここに取り込む。

産業上の利用可能性

[0189] 本発明は、例えば、企業が自社商品の風評被害がウェブ (Web) 上で将来的に起こらないかどうかを監視するためのシステム等に利用可能である。

符号の説明

- [0190] 1 1、1 2 データ入力手段
- 1 2 A 学習用データ入力部
- 1 2 B 属性情報入力部
- 1 2 C 予測用データ入力部
- 2 1 データ記憶手段
- 2 1 A 学習データ記憶部
- 2 1 B 学習処理情報保存部
- 3 1、3 2 データ集中予測手段
- 4 1、4 2 学習データ抽出処理部
- 4 1 A、4 2 A 学習データ記憶処理機能
- 5 1、5 2 学習処理部
- 5 1 A データ更新処理機能
- 5 1 B 再学習処理機能
- 5 2 A 学習分類機能
- 5 2 B 影響力処理機能
- 6 1、6 2 情報集計部
- 6 1 A、6 2 A 学習データ集計機能
- 6 1 B、6 2 B 予測データ集計機能
- 6 2 C グループ化機能
- 6 3 グループ集計機能
- 7 1、7 2 予測処理部
- 7 1 A 予測結果出力機能
- 7 2 A 予測分類機能
- 8 1、8 2 データ集中予測装置

請求の範囲

[請求項1] 複数のノードから発信されるデータを対応する属性情報と共に受信し時系列データとして取り込むデータ入力手段と、この取り込んだ時系列データを学習データとして記憶するデータ記憶手段と、前記記憶した時系列データのデータ構造を分析し後発するデータ集中を予測するデータ集中予測手段と、を備えたデータ集中予測装置であって、

前記データ集中予測手段は、前記データ入力手段が取り込む時系列データを予め設定された単位時刻ごとに経時的に一時記憶処理すると共に、この一時記憶処理データを利用して各時系列データの入力時刻の直前の時刻を基準とした過去の一定期間内の時系列データに基づいて予め設定される変動許容範囲を逸脱する当該時系列データを前記後発するデータ集中の予測に必要な追加学習データとして連続的に抽出処理する学習データ抽出処理部を備え、

この学習データ抽出処理部が、前記連続的に抽出した追加学習データを前記データ記憶手段に一括して記憶処理する学習データ記憶処理機能を備えたことを特徴とするデータ集中予測装置。

[請求項2] 前記請求項1に記載のデータ集中予測装置において、

前記追加学習データの抽出に際して設定される変動許容範囲を、前記過去の一定期間内の時系列データが内包する前記データ集中の予測にかかる属性のデータの平均値と分散に基づいて前記学習データ抽出処理部が算出し設定することを特徴としたデータ集中予測装置。

[請求項3] 前記請求項2に記載のデータ集中予測装置において、

前記データ集中予測手段が、

前記データ記憶手段に一括記憶された追加学習データが内包する前記データ集中の予測にかかる属性のデータを前記単位時刻に対応づけて学習集計データとして集計する学習データ集計機能を有する情報集計部と、

前記データ集中にかかる予測値に対する各ノードの影響力を示す影

響力データを前記学習集計データとの関係で算出すると共に、この影響力データ及び当該算出に用いた学習集計データを前記データ記憶手段に保存処理する学習処理部と、を備えたことを特徴とするデータ集中予測装置。

[請求項4]

前記請求項3に記載のデータ集中予測装置において、

前記情報集計部は、更に、予め設定された時間間隔で発せられる予測要求に応じて前記データ入力手段が取り込んだ時系列データに含まれる前記予測にかかる属性のデータを前記単位時刻に対応づけて予測集計データとして集計する予測データ集計機能を備え、

前記データ集中予測手段が、更に、前記予測集計データと前記影響力データとに基づいて前記予測値を算出処理する予測処理部を備えたことを特徴とするデータ集中予測装置。

[請求項5]

前記請求項4に記載のデータ集中予測装置において、

前記学習処理部は、更に、リアルタイムでの前記影響力データの算出に際して前記データ記憶手段に予め影響力データ及び学習集計データが保存されていた場合には、リアルタイムにて算出した影響力データ及び当該算出に用いた学習集計データにより前記データ記憶手段内の保存情報を更新するデータ更新処理機能を備えたことを特徴とするデータ集中予測装置。

[請求項6]

前記請求項5に記載のデータ集中予測装置において、

前記学習処理部は、更に、リアルタイムでの前記影響力データの算出に際して前記データ記憶手段に予め学習集計データが保存されていた場合には、その保存された学習集計データとリアルタイムにて前記情報集計部から取得する学習集計データとを結合処理すると共にこれを使用して前記影響力データを算出する再学習処理機能を備えたことを特徴するデータ集中予測装置。

[請求項7]

前記請求項4に記載のデータ集中予測装置において、

前記情報集計部は、

前記各ノードの属性データの共通性に基づいてグループを定めると共に各ノードを1以上のグループに帰属させてグループ情報を作成するグループ化機能と、

このグループ情報を前記予測にかかる属性のデータに代えて、前記時刻に対応づけることにより前記学習集計データ又は前記予測集計データを作成するグループ集計機能と、を備えたことを特徴とするデータ集中予測装置。

[請求項8] 前記請求項7に記載のデータ集中予測装置において、

前記学習処理部は、前記各ノードにかかる影響力データを算出するに際して、前記データ集中に対する各グループの影響力を前記学習集計データとの関係で算出すると共に、前記各ノードについて当該ノードの属するグループの影響力の加算値を当該ノードの影響力とする影響力処理機能を備えたことを特徴とするデータ集中予測装置。

[請求項9] 複数のノードから発信されるデータを対応する属性情報と共に受信し時系列データとして取り込むデータ入力手段と、この取り込んだ時系列データを学習データとして記憶するデータ記憶手段と、前記記憶した時系列データのデータ構造を分析し後発するデータ集中を予測するデータ集中予測手段と、を備えたデータ集中予測装置にあって、

前記データ集中予測手段は、前記データ集中の予測用の時系列データを抽出し処理する学習データ抽出処理部を備えており、

前記データ入力手段にて取り込まれた時系列データを予め設定された単位時刻ごとに経時的に一時記憶処理し、

この一時記憶処理データを利用して各時系列データが各々の入力時刻の直前の時刻を基準とした過去の一定期間内の時系列データに基づいて設定される変動許容範囲を逸脱するものであるか否かを判定し、

前記変動許容範囲を逸脱する旨判定した際に、当該判定にかかる時系列データを前記後発するデータ集中の予測に必要な追加学習データとして連続的に抽出し、

この連続的に抽出した追加学習データを前記データ記憶手段に一括して記憶処理すると共に、

これら一連の各工程内容を前記学習データ抽出処理部が順次実行し、

前記データ記憶手段に一括記憶された追加学習データ及び既存の学習データにより特定される前記時系列データのデータ構造に基づいて、前記データ集中予測手段の予測処理部が前記データ集中を予測することを特徴としたデータ集中予測方法。

[請求項10]

複数のノードから発信されるデータを対応する属性情報と共に受信し時系列データとして取り込むデータ入力手段と、この取り込んだ時系列データを学習データとして記憶するデータ記憶手段と、前記記憶した時系列データのデータ構造を分析し後発するデータ集中を予測するデータ集中予測手段と、を備えたデータ集中予測装置にあって、

前記データ入力手段にて取り込まれた時系列データを予め設定された単位時刻ごとに経時的に一時記憶処理するデータ一時記憶処理機能、

この一時記憶処理データを利用して各時系列データが各々の入力時刻の直前の時刻を基準とした過去の一定期間内の時系列データに基づいて設定される変動許容範囲を逸脱するものであるか否かを判定する変動許容判定機能、

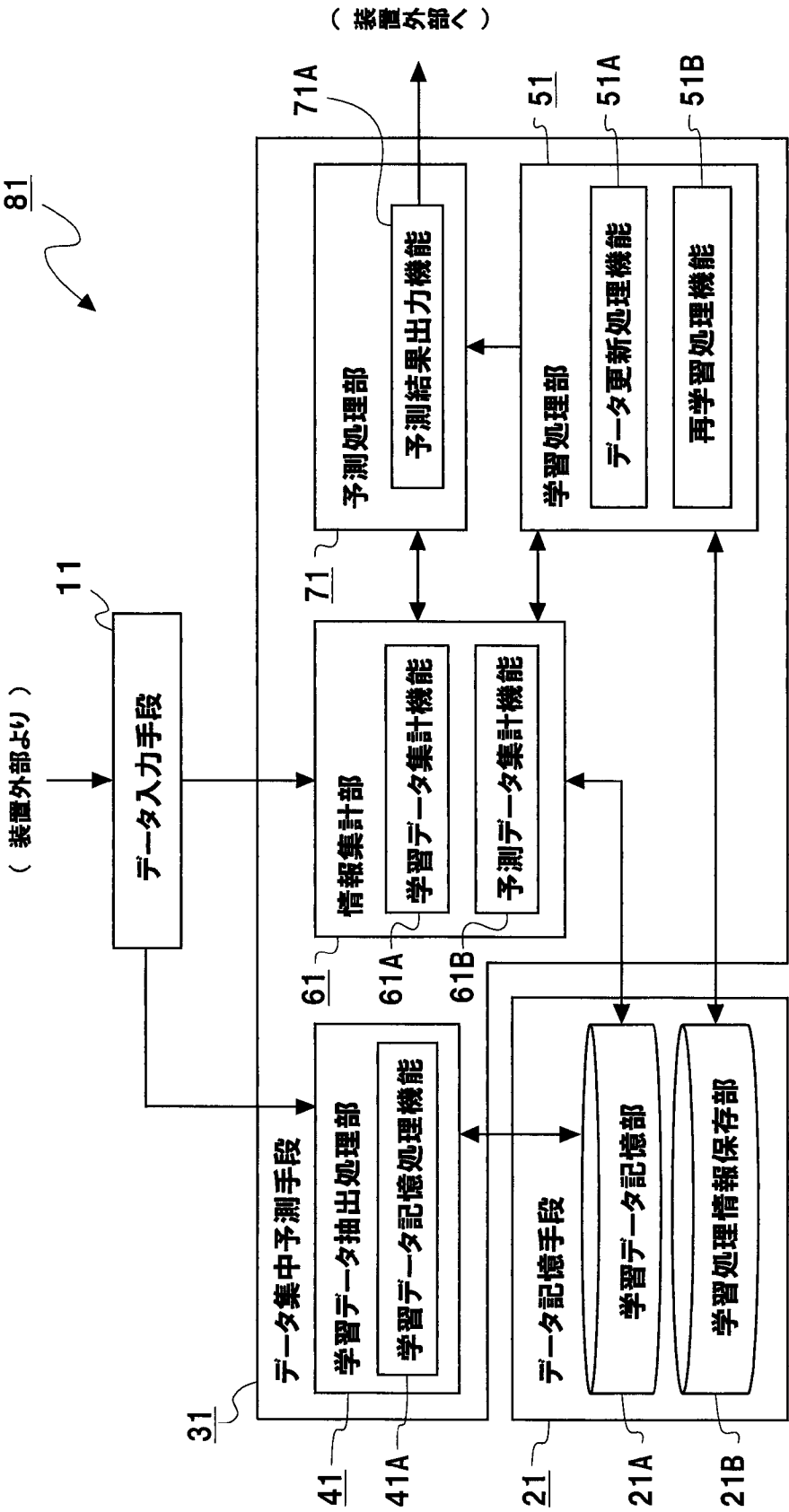
前記変動許容範囲を逸脱する旨判定した際に、当該判定にかかる時系列データを前記後発するデータ集中の予測に必要な追加学習データとして連続的に抽出する学習データ抽出機能、

およびこの連続的に抽出した追加学習データを前記データ記憶手段に一括して記憶処理する学習データ記憶処理機能、

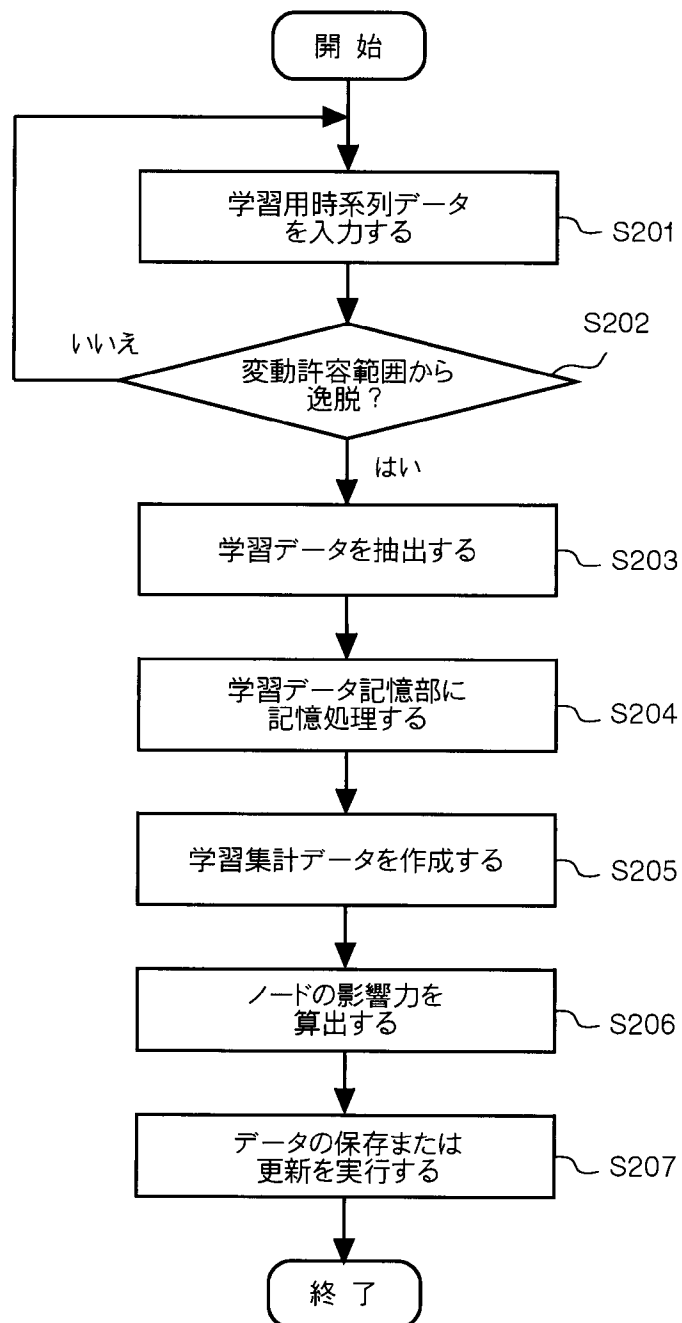
この一括記憶された追加学習データ及び既存の学習データにより特定される前記時系列データのデータ構造に基づいて前記データ集中を予測する予測処理機能を設け、

これら各情報処理機能を、前記データ集中予測手段が備えているコンピュータに実現させるようにしたことを特徴とするデータ集中予測プログラム。

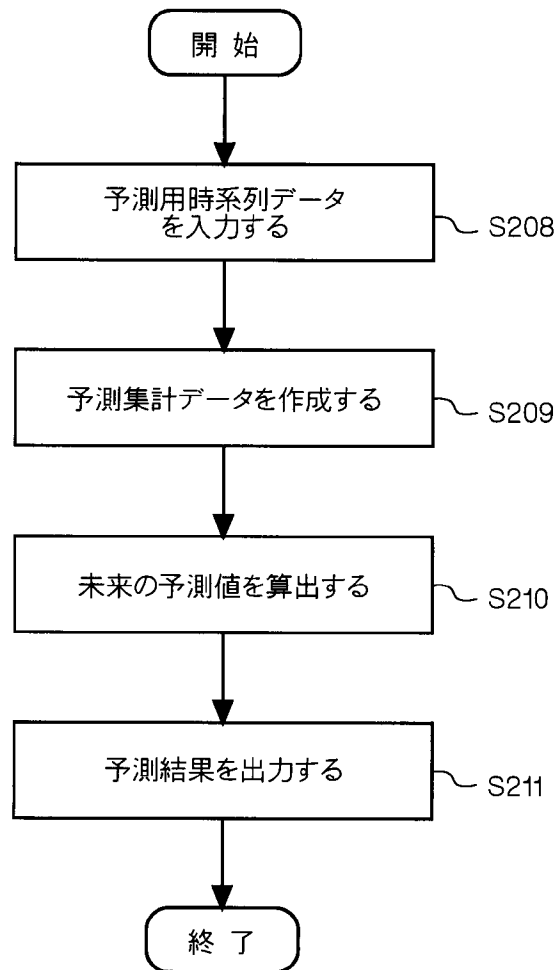
[図1]



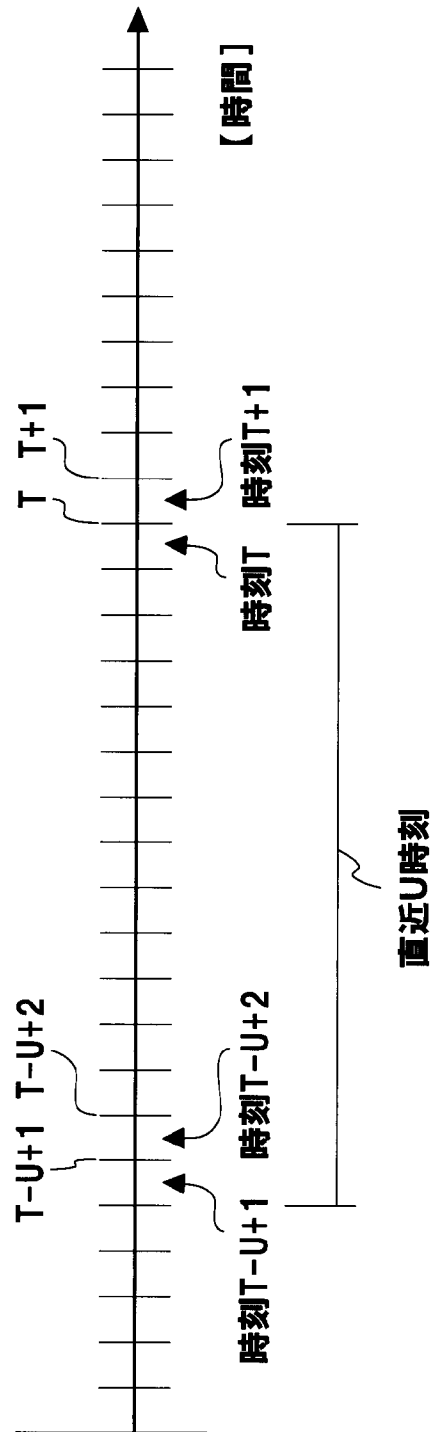
[図2]



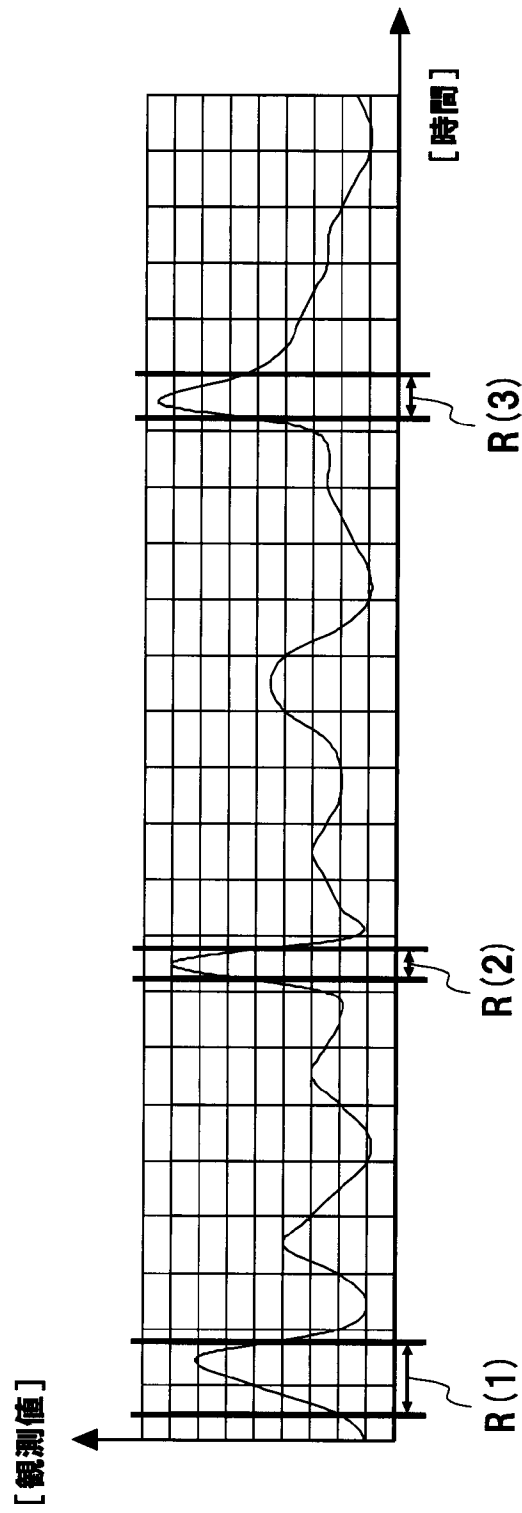
[図3]



[図4]



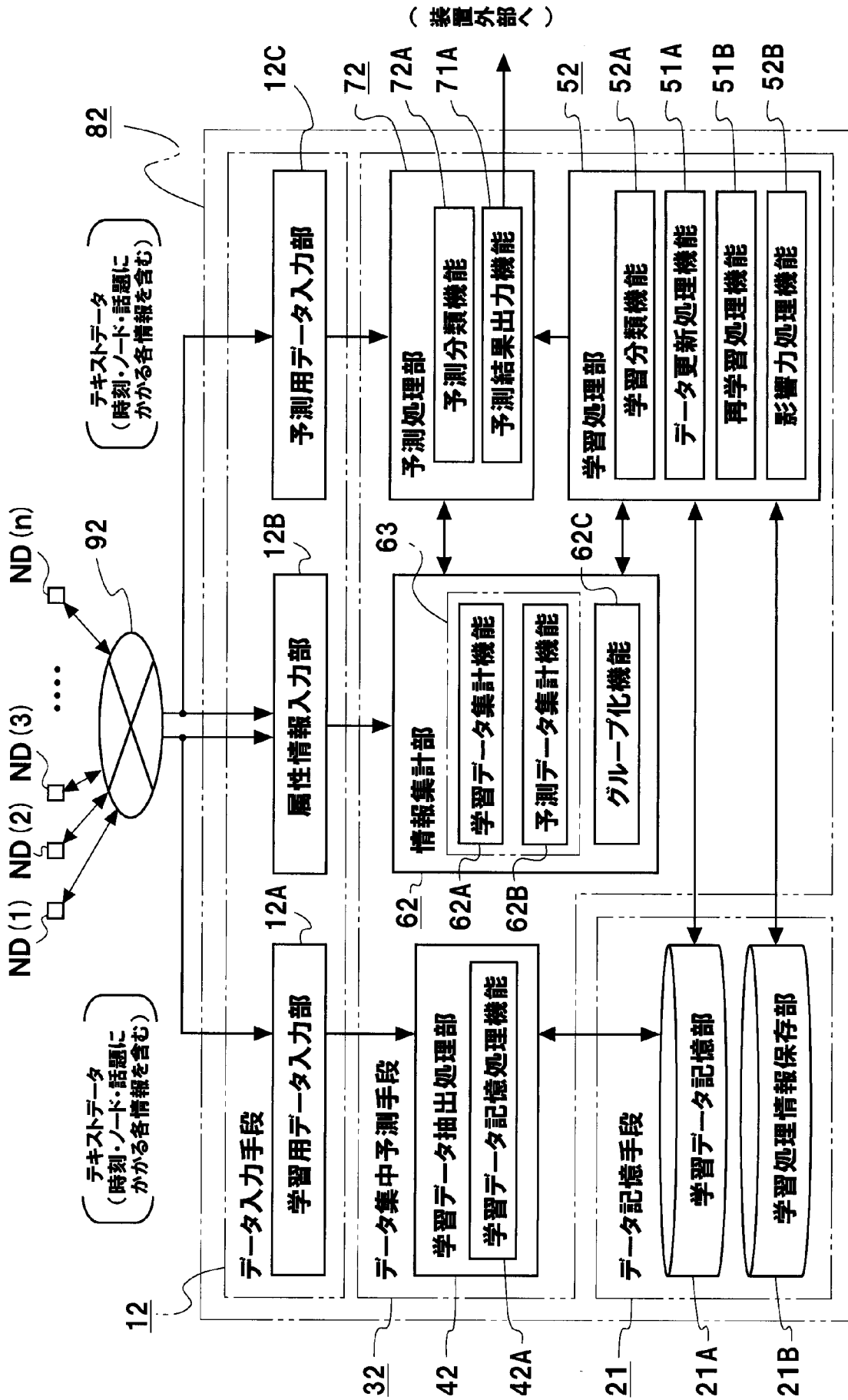
[図5]



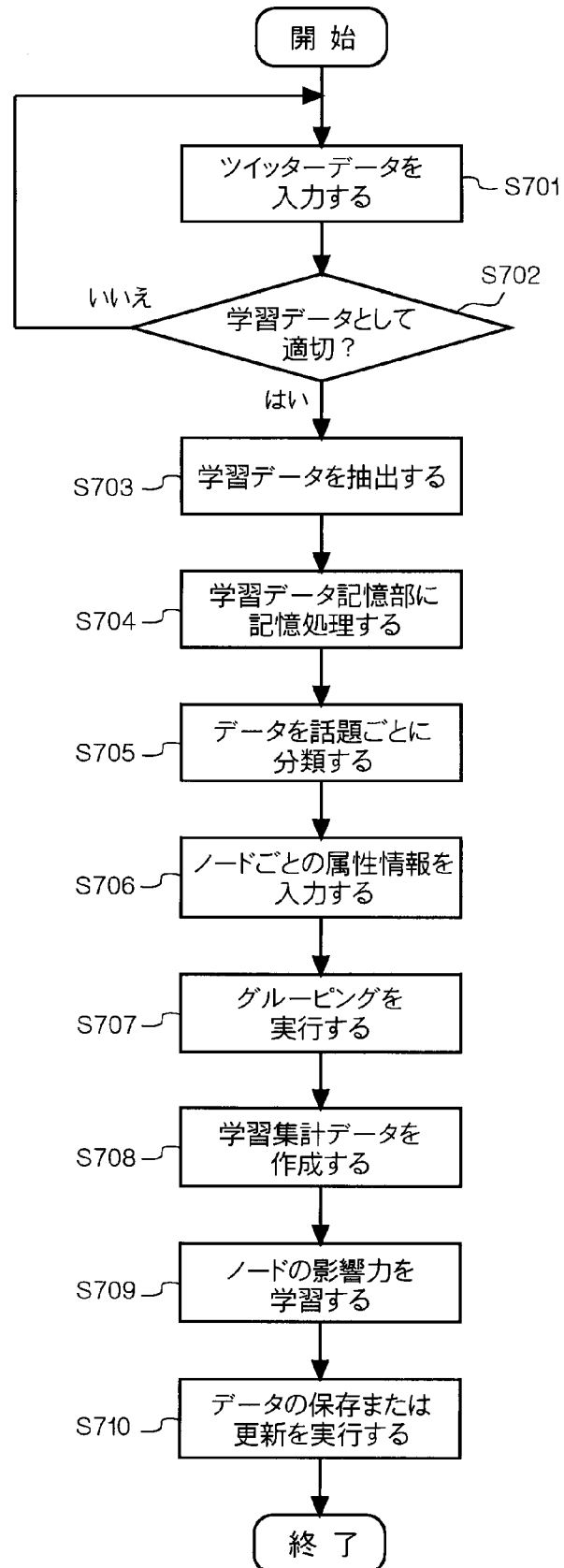
←→ : 学習データとして有効と判定される期間 (有効学習期間)

縦軸 : 観測値(データ集中の予測にかかる属性のデータ) / 横軸 : 時間(時刻)

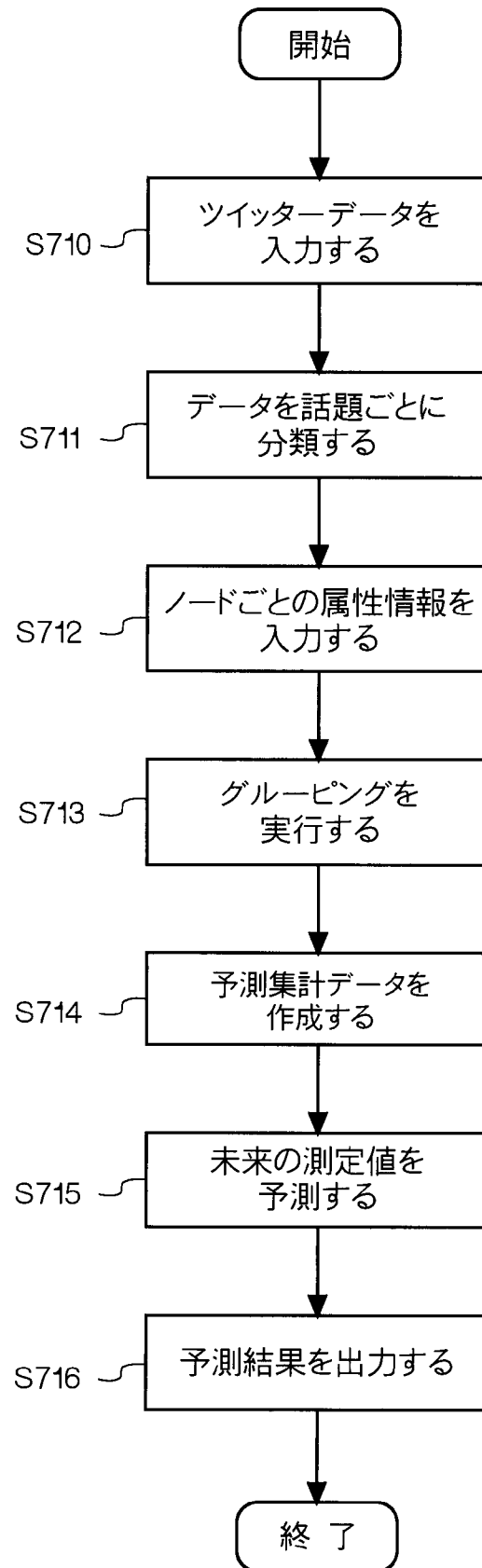
[図6]



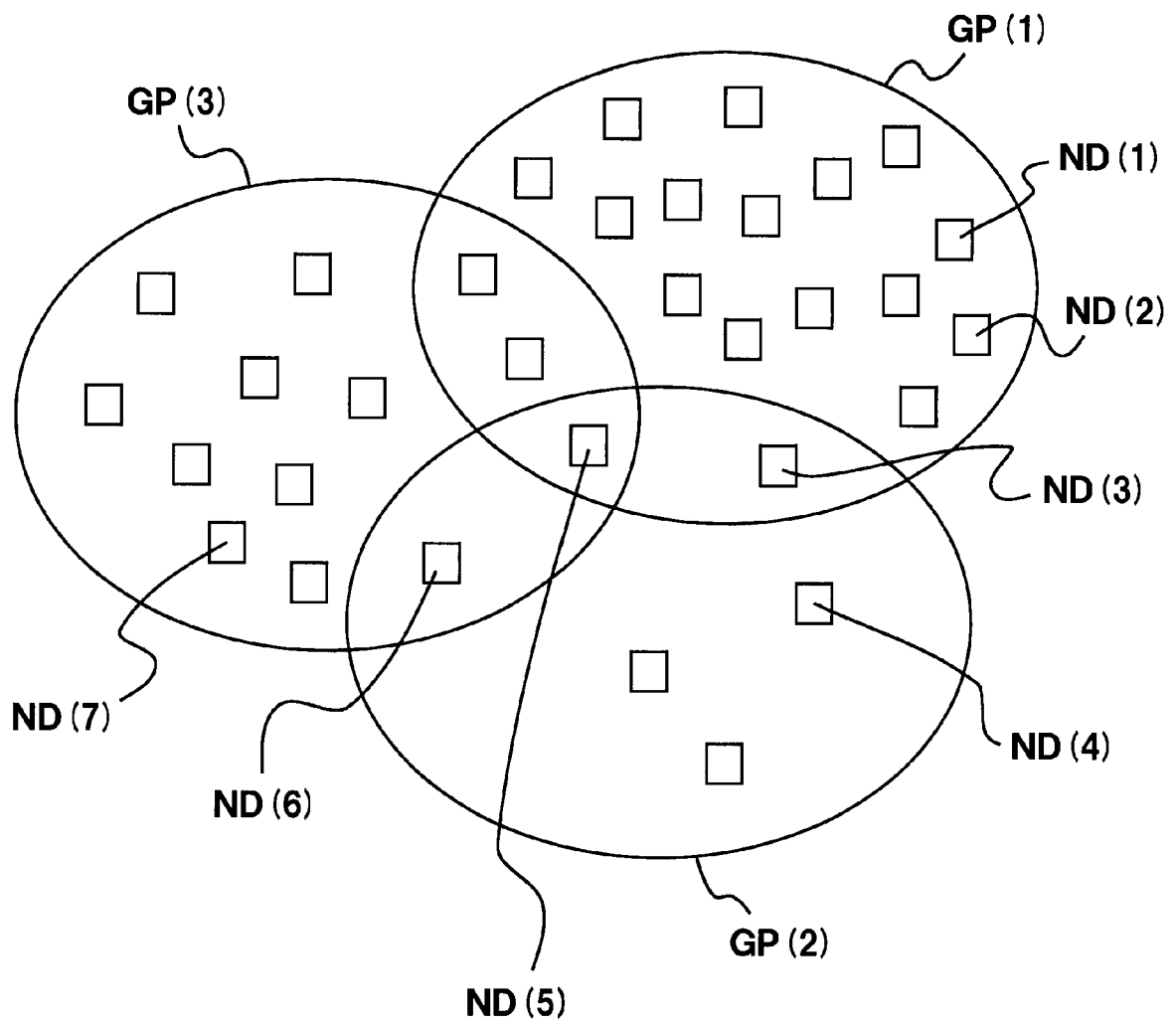
[図7]



[図8]



[図9]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2013/074367

A. CLASSIFICATION OF SUBJECT MATTER

G06N3/00(2006.01)i, G06F19/00(2011.01)i, G06N5/04(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N3/00, G06F19/00, G06N5/04

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2013
Kokai Jitsuyo Shinan Koho	1971-2013	Toroku Jitsuyo Shinan Koho	1994-2013

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Toshimitsu TAKAHASHI et al., "Topic Emergence Detection in Social Network Using Probabilistic Models of Links", IEICE Technical Report, 21 March 2011 (21.03.2011), vol.110, no.476, pages 169 to 176	1-10
A	Yutaka HIROKAWA et al., "Traffic Leap Detection for Internet Backbone", IEICE Technical Report, 01 March 2007 (01.03.2007), vol.106, no.577, pages 403 to 408	1-10

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
15 October, 2013 (15.10.13)

Date of mailing of the international search report
22 October, 2013 (22.10.13)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. G06N3/00(2006.01)i, G06F19/00(2011.01)i, G06N5/04(2006.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（I P C））

Int.Cl. G06N3/00, G06F19/00, G06N5/04

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 1 3 年
日本国実用新案登録公報	1 9 9 6 - 2 0 1 3 年
日本国登録実用新案公報	1 9 9 4 - 2 0 1 3 年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	高橋 俊充ほか2名, リンクの確率構造を用いたソーシャルネットワークにおける話題拡大検出, 電子情報通信学会技術研究報告, 2011.03.21, V o l . 1 1 0, N o . 4 7 6, p p . 1 6 9 - 1 7 6	1 - 1 0
A	廣川 裕ほか3名, インターネットバックボーンにおけるトラヒックリープ検出方法, 電子情報通信学会技術研究報告, 2007.03.01, V o l . 1 0 6, N o . 5 7 7, p p . 4 0 3 - 4 0 8	1 - 1 0

☐ C欄の続きにも文献が列挙されている。☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの

「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「O」口頭による開示、使用、展示等に言及する文献

「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」同一パテントファミリー文献

国際調査を完了した日

1 5 . 1 0 . 2 0 1 3

国際調査報告の発送日

2 2 . 1 0 . 2 0 1 3

国際調査機関の名称及びあて先

日本国特許庁（I S A / J P）

郵便番号100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

塚田 肇

5 B

3 6 5 2

電話番号 03-3581-1101 内線 3545