(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2014-120140
(P2014-120140A)
(43)公開日　平成26年6月30日(2014.6.30)

(51)Int.Cl.	F I	テーマコード（参考）
G06F 17/30 (2006.01)	G06F 17/30 210D	

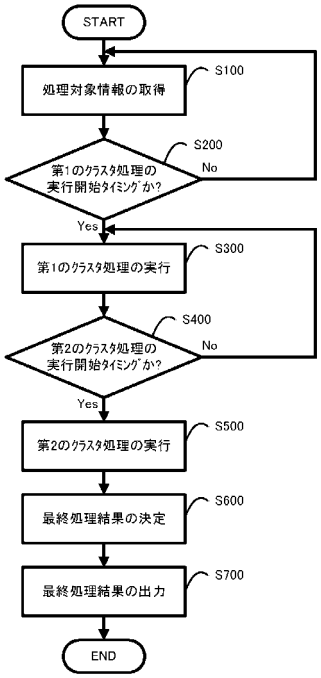
審査請求 未請求　請求項の数 8　ＯＬ　（全 22 頁）

(21)出願番号　特願2012-277491 (P2012-277491)
(22)出願日　平成24年12月19日 (2012.12.19)

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100108187
弁理士　横山　淳一
(72)発明者　岩倉　友哉
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

(54)【発明の名称】クラスタ処理方法、クラスタ処理装置およびプログラム

(57)【要約】
【課題】開示の技術は、高速なクラスタ処理を実現可能とする。
【解決手段】本開示の技術における解決手段の一観点は、第２のクラスタ手法と異なりかつ該第２のクラスタ手法よりも高速な第１のクラスタ手法によって対象情報に対して第１の分類処理を実行し、前記第１の分類処理の結果に基づいて前記第２のクラスタ手法によって第２の分類処理を実行し、前記第１の分類処理の結果と前記第２の分類処理の結果とに基づいて前記対象情報の分類を決定する、クラスタ処理方法、クラスタ処理装置およびプログラムである。
【選択図】図４



【特許請求の範囲】
【請求項１】
コンピュータが、
　第２のクラスタ手法と異なりかつ該第２のクラスタ手法よりも高速な第１のクラスタ手法によって対象情報に対して第１の分類処理を実行し、
　前記第１の分類処理の結果に基づいて前記第２のクラスタ手法によって第２の分類処理を実行し、
　前記第１の分類処理の結果と前記第２の分類処理の結果とに基づいて前記対象情報の分類を決定する、
ことを特徴とするクラスタ処理方法。
【請求項２】
　前記対象情報の分類を決定することは、前記第１の分類処理の結果と前記第２の分類処理の結果間で分類の変更がなかった前記第１の分類処理の結果または前記第２の分類処理の結果に基づいて前記対象情報の分類を決定する、
ことを特徴とする請求項１記載のクラスタ処理方法。
【請求項３】
　前記第１の分類処理の結果は、前記対象情報に関連した分類個数情報を含み、
　前記第２の分類処理は、前記分類個数情報を前記第２のクラスタ手法における初期情報として実行される、
ことを特徴とする請求項１または２記載のクラスタ処理方法。
【請求項４】
　前記第１の分類処理の結果は、前記対象情報に含まれるキーワードを基にした前記対象情報の特徴情報を含み、
　前記第２の分類処理は、前記対象情報の特徴情報を前記第２のクラスタ手法における前記初期情報として実行される、
ことを特徴とする請求項３記載のクラスタ処理方法。
【請求項５】
　前記第１の分類処理は、前記対象情報を用いた前記キーワード別の集合を作成し、異なる前記キーワードに対応する前記集合を異なる分類とする、
ことを特徴とする請求項４記載のクラスタ処理方法。
【請求項６】
　前記第２のクラスタ手法は、処理対象情報の類似性を基に処理対象情報を分類する手法である、
ことを特徴とする請求項１乃至５のいずれか１項に記載のクラスタ処理方法。
【請求項７】
　第２のクラスタ手法と異なりかつ該第２のクラスタ手法よりも高速な第１のクラスタ手法によって対象情報に対して第１の分類処理を実行する第１処理部と、
　前記第１の分類処理の結果に基づいて前記第２のクラスタ手法によって第２の分類処理を実行する第２処理部と、
　前記第１の分類処理の結果と前記第２の分類処理の結果とに基づいて前記対象情報の分類を決定する決定部と、
を有することを特徴とするクラスタ処理装置。
【請求項８】
コンピュータに、
　第２のクラスタ手法と異なりかつ該第２のクラスタ手法よりも高速な第１のクラスタ手法によって対象情報に対して第１の分類処理を実行し、
　前記第１の分類処理の結果に基づいて前記第２のクラスタ手法によって第２の分類処理を実行し、
　前記第１の分類処理の結果と前記第２の分類処理の結果とに基づいて前記対象情報の分類を決定する、

処理を実行させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、クラスタ処理方法、クラスタ処理装置およびプログラムに関する。
【背景技術】
【０００２】
クラスタリング技術に関する先行技術として、文書類似ベクトルを用いて対象文書とクラスタ重心との距離を算出し、さらに同一の対象文書に対して一回目の分類に利用した文書類似ベクトルの次元数を増加させて二回目の分類を行い、安定クラスタの文書を対象から除いて次の対象文書を選定して分類試行を繰り返す技術が知られている。（例えば、特許文献１参照）
また、他の先行技術として、データ集合を部分クラスタの集合に変換し（大分類し）、部分クラスタの集合をクラスタリングするにあたり部分クラスタの局所的な密度に関する属性を考慮して詳細分類を行う技術が知られている。（例えば、特許文献２参照）
【先行技術文献】
【特許文献】
【０００３】
【特許文献１】特開２００２－１８３１７１号公報
【特許文献２】特開２０１０－１３４６３２号公報
【発明の概要】
【発明が解決しようとする課題】
【０００４】
前記先行技術は、不安定なクラスタの文書や部分クラスタを詳細分類することで、高精度なクラスタリング結果を得ることができる。
【０００５】
しかしながら、前記先行技術では、このような詳細分類が繰り返されるためにクラスタ処理に多くの時間を要するという問題がある。特に、大規模データを処理対象にすると、そのデータ量の多さから必然的に不安定なクラスタの数も多くなる。そのため、前記先行技術により大規模データを対象にしたクラスタ処理においては、更に多くの処理時間が必要となる。
【０００６】
本願は、高速なクラスタ処理を実現可能とするクラスタ処理方法、クラスタ処理装置およびプログラムを提供することを目的とする。
【課題を解決するための手段】
【０００７】
上記課題を解決するために、本実施例に開示のクラスタ処理方法は、コンピュータが、第２のクラスタ手法と異なりかつ該第２のクラスタ手法よりも高速な第１のクラスタ手法によって対象情報に対して第１の分類処理を実行し、前記第１の分類処理の結果に基づいて前記第２のクラスタ手法によって第２の分類処理を実行し、前記第１の分類処理の結果と前記第２の分類処理の結果とに基づいて前記対象情報の分類を決定する。
【発明の効果】
【０００８】
本実施例の一観点によれば、高速なクラスタ処理が実現される。
【図面の簡単な説明】
【０００９】
【図１】コンピュータシステムを示す。
【図２】コンピュータのハードウェアを示す。
【図３】コンピュータの機能ブロックを示す。
【図４】処理全体のフローチャートを示す。

【図５】処理例を示す。
【図６】取得処理の一例のフローチャートを示す。
【図７】第１クラスタ処理の一例のフローチャートを示す。
【図８】第１クラスタ処理におけるキーワード抽出処理の一例のフローチャートを示す。
【図９】第１クラスタ処理における集合作成処理の一例のフローチャートを示す。
【図１０】第１クラスタ処理におけるベクトル情報作成処理の一例のフローチャートを示す。
【図１１】第２クラスタ処理の一例のフローチャートを示す。
【図１２】決定処理の一例のフローチャートを示す。
【発明を実施するための形態】
【００１０】
以下、図面を参照して開示の技術の実施形態の一例を詳細に説明する。
【００１１】
図１は、コンピュータシステム１を示す。コンピュータシステム１は、例えば、コンピュータ１００、ネットワーク２００、サーバ３００、ストレージシステム３２０、コンピュータ４００、コンピュータ５００、ＮＡＳ（Ｎｅｔｗｏｒｋ Ａｔｔａｃｈｅｄ Ｓｔｏｒａｇｅ）６００を含む。本例において、コンピュータ１００、サーバ３００、コンピュータ４００、コンピュータ５００、ＮＡＳ６００は、ネットワーク２００にそれぞれ接続されている。
【００１２】
コンピュータ１００は、その詳細を後述する装置であり、クラスタ処理を実行する。コンピュータ１００は、例えば、サーバ、ワークステーション、パーソナルコンピュータ、インターネットアプライアンス、ゲーム機などである。コンピュータ１００として、クラスタ処理にかかる処理負荷やその処理結果の提供形態などに応じて、適宜の装置が選択されればよい。
【００１３】
ネットワーク２００は、ＬＡＮやインターネット等であり、それに接続される装置間でのデータ通信を可能とする。
【００１４】
サーバ３００は、ソーシャルメディアを実現する装置である。ソーシャルメディアは、例えば、電子掲示板、ブログ、ウィキ、ツイッター、ポッドキャスト、ソーシャルブックマーク、 ソーシャル・ネットワーキング・サービス、画像や動画の共有サイト、通販サイトのカスタマーレビューなどである。サーバ３００は、ネットワーク２００を介して受信したデータ登録要求やデータ送信要求に応じて、それら要求で示される処理を実行し、また、それら要求を発行した他の装置に要求対象データの送信等を実行する。サーバ３００は、ソーシャルメディアを実現するためのデータを格納するストレージシステム３２０を備えてもよい。ストレージシステム３２０は、例えば、ＲＡＩＤ（Ｒｅｄｕｎｄａｎｔ Ａｒｒａｙｓ ｏｆ Ｉｎｅｘｐｅｎｓｉｖｅ Ｄｉｓｋｓ）システムである。サーバ３００は、ソーシャルメディア上の大規模データを扱う。なお、サーバ３００は、ワークステーション、パーソナルコンピュータなど他のコンピュータであってもよい。
【００１５】
コンピュータ１００は、上記のようにサーバ３００が提供するソーシャルメディア上の大規模データを対象にクラスタ処理を実行し、ソーシャルメディア上で盛り上がっている話題（内容やキーワード）を抽出する。例えば、コンピュータ１００は、ソーシャルメディアの情報である例えばテキストをその内容の類似性によってまとめあげることによって、ソーシャルメディア上の情報の中から盛り上がっている話題ごとに情報をクラスタリングする。
【００１６】
コンピュータ４００は、サーバ３００により実現されるソーシャルメディアを利用し、そのサービスを受ける装置である。例えば、コンピュータ４００がネットワーク２００を

介してサーバ３００に前述のデータ登録要求やデータ送信要求を発行する。そのような要求に対して、サーバ３００は、要求にて示される処理を実行する。

【００１７】

コンピュータ５００は、例えば、コンピュータ１００により実行されたクラスタ処理の処理結果を取得して表示する。また、コンピュータ５００は、コンピュータ１００の動作設定を管理する装置であってもよい。

【００１８】

ＮＡＳ６００は、ネットワーク２００に接続されたストレージシステムである。例えば、前述のサーバ３００は、直接または他のネットワークを介して間接にネットワーク２００に接続されたＮＡＳ６００をデータの格納先として利用してもよい。この場合、サーバ３００は、ネットワーク２００を介して、データのライト要求やリード要求をＮＡＳ６００に発行することになる。ＮＡＳ６００は、他装置、例えばサーバ３００からのこのような要求を受信し、その要求で示される処理を実行する。ＮＡＳ６００は、その要求がデータのライトを示す場合は指定されたデータを自身が有する記憶装置（例えばハードディスク）に記録し、要求がデータのリードを示す場合は指定されたデータを記憶装置から読み出して要求の発行元装置にそれを送信する。ＮＡＳ５００は、ＲＡＩＤシステムであってもよい。

【００１９】

図２は、コンピュータ１００のハードウェアを示す。この図２に示されるハードウェアは、コンピュータ１００を構成するハードウェアの一例であり、少なくとも本実施形態に記載される処理の実行に必要なハードウェア構成を備えていればよい。

【００２０】

図２に示されるコンピュータ１００は、例えば、プロセッサ１０、ＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）２０、ドライブ装置３０、記憶媒体３２、入力インターフェース（Ｉ／Ｆ）４０、入力デバイス４２、出力インターフェース（Ｉ／Ｆ）５０、出力デバイス５２、通信インターフェース（Ｉ／Ｆ）６０およびバス７０などを含む。それぞれのハードウェア構成は、バス７０を介して接続されている。

【００２１】

プロセッサ１０は、例えば、ＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）、ＭＰＵ（Ｍｉｃｒｏ－Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）、ＤＳＰ（Ｄｉｇｉｔａｌ Ｓｉｇｎａｌ Ｐｒｏｃｅｓｓｏｒ）などの処理回路である。

【００２２】

ＲＡＭ２０は読み書き可能なメモリ装置であって、例えば、ＳＲＡＭ（Ｓｔａｔｉｃ ＲＡＭ）やＤＲＡＭ（Ｄｙｎａｍｉｃ ＲＡＭ）などの半導体メモリである。なお、ＲＡＭではなく、フラッシュメモリなどであってもよい。

【００２３】

ドライブ装置３０は、記憶媒体３２にアクセスする。記憶媒体３２は、データを記憶している。ドライブ装置３０は、記憶媒体３２へのデータのライト、記憶媒体３２からのデータのリードの少なくともいずれか一方を行う。記憶媒体３２は、例えば、ハードディスク、ＳＳＤ（Ｓｏｌｉｄ Ｓｔａｔｅ Ｄｒｉｖｅ）などのフラッシュメモリ、ＣＤ（Ｃｏｍｐａｃｔ Ｄｉｓｃ）、ＤＶＤ（Ｄｉｇｉｔａｌ Ｖｅｒｓａｔｉｌｅ Ｄｉｓｃ）、ブルーレイディスクなどである。コンピュータ１００は、記憶媒体３２の種類に応じたドライブ装置３０を備えればよい。

【００２４】

入力インターフェース４０は、入力デバイス４２が接続され、入力デバイス４２から受信した入力信号をプロセッサ１０に伝達する回路である。入力デバイス４２は、利用者による操作に応じた入力信号を出力する装置である。入力デバイス４２は、例えば、キーボードやコンピュータ１００に設置されたボタンなどのキー装置や、マウスやタッチパネルなどのポインティングデバイスである。

【００２５】

出力インターフェース５０は、出力デバイス５２が接続され、出力デバイス５２に、プロセッサ１００の指示に応じた出力を実行させる回路である。出力デバイス５２は、コンピュータ１００の制御に応じて情報を出力する装置である。出力デバイス５２は、例えば、ディスプレイなどの画像出力装置（表示デバイス）や、スピーカーなどの音声出力装置、プリンタなどである。また、例えば、タッチスクリーンなどの入出力装置が、入力デバイス４２および出力デバイス５２として用いられる。また、入力デバイス４２及び出力デバイス５２は、コンピュータ１００と一体であってもよいし、コンピュータ１００に含まれず、例えば、コンピュータ１００に外部から接続する装置であってもよい。また、例えば、コンピュータ１００が通常動作時にこのコンピュータ１００を利用するユーザに対して情報提供するための出力デバイス５２を必要としないサーバのようなコンピュータであれば、入力デバイス４２および出力デバイス５２を不要とする構成としてもよい。また、そのような場合、入力インターフェース４０および出力インターフェース５０も不要とする構成としてもよい。

【００２６】

通信インターフェース６０は、ネットワーク２００に接続され、ネットワーク２００を介した通信の制御を行なう回路である。通信インターフェース６０は、例えば、Ｅｔｈｅｒｎｅｔ（登録商標）カードに代表されるネットワークインターフェースカード（ＮＩＣ）などである。

【００２７】

例えば、プロセッサ１０は、記憶媒体３２からプログラムをリードし、それをＲＡＭ２０にロードする。また、プロセッサ１０は、プログラムで利用されるデータをＲＡＭ２０にロードする。このように、ＲＡＭ２０は、プロセッサ１０によるプログラム実行のためのワークエリアとして用いられる。プログラムは、オペレーティング・システム（ＯＳ）やアプリケーションプログラムなどであり、所定の処理手順を実行する命令を含む。例えば、このようなプログラムの動作により、本実施形態の処理機能がコンピュータ１００上に実現される。

【００２８】

図３は、コンピュータ１００の機能ブロックを示す。各機能ブロックは、前述のプログラムの動作によってプロセッサ１０により実現される。

【００２９】

コンピュータ１００上で実現される機能ブロックは、例えば、制御部１１０、記憶部１２０、取得部１３０、第１クラスタ処理部１４０、第２クラスタ処理部１５０、決定部１６０および出力部１７０を含む。

【００３０】

制御部１１０は、本実施形態の処理機能を実現するため、各機能ブロックを制御する。記憶部１２０は、各種データを記憶する。なお、記憶部１２０は、その機能を実現するため、図２におけるＲＡＭ２０または記憶媒体３２も含む。

【００３１】

取得部１３０は、処理対象となるデータ、すなわち、本実施形態においてはソーシャルメディア上の情報を取得する。以後、取得部１３０によって取得された情報を処理対象情報と呼ぶことがある。また、取得部１３０は、取得した処理対象情報を記憶部１２０に格納する。本実施形態において取得部１３０が取得するソーシャルメディア上の情報は、例えばテキスト情報である。また、処理対象情報は、複数のファイルであってもいいし、複数個の文章を含む１つのファイルであってもよい。取得部１３０は、図２における通信インターフェース６０を含んでもよい。

【００３２】

第１クラスタ処理部１４０は、記憶部１２０に格納された処理対象情報を対象に、第１のクラスタ処理を実行する。本実施形態において、第１のクラスタ処理は、例えば、キーワード抽出処理である。このキーワード抽出処理は、処理対象情報に含まれるキーワードを抽出する。抽出されたキーワードは、処理対象情報の分類処理に利用される。この処理

の詳細は後述する。第１クラスタ処理部１４０は、第１のクラスタ処理の結果を記憶部１２０に格納する。また、第１クラスタ処理部１４０は、記憶部１２０をワークスペースとして、各種データを記憶部１２０に一時格納し、それらデータを利用して第１のクラスタ処理を実行する。

【００３３】

第２クラスタ処理部１５０は、記憶部１２０に格納されている第１クラスタ処理部１４０による処理結果を入力情報および初期設定情報として使用し、処理対象情報を対象にした第２のクラスタ処理を実行する。本実施形態において、第２のクラスタ処理は、例えば、Ｋ－ｍｅａｎｓ（Ｋ－平均法）である。このＫ－ｍｅａｎｓは、処理対象となる情報の内容を考慮し、類似する情報を分類（クラスタリング）する。また、後述するように、一般的なＫ－ｍｅａｎｓでは、最終解を得るために設定情報を何回か変更して処理試行を行う。これらが理由で、Ｋ－ｍｅａｎｓは、最終の処理結果を得るまでに処理時間がかかる。

【００３４】

これに対し、第１クラスタ処理部１４０が実行する前述の第１のクラスタ処理の一例として示したキーワード抽出処理は、Ｋ－ｍｅａｎｓのように処理対象となる情報間の類似性を考慮せず、情報内に含まれるキーワードをピックアップし、キーワードとそれを含む情報との対応付けにより分類処理を行う。このため、同じ情報の分類処理において、キーワード抽出処理を一例にする第１のクラスタ処理は、Ｋ－ｍｅａｎｓを一例にする第２のクラスタ処理より高速である。

【００３５】

第２クラスタ処理部１５０は、第２のクラスタ処理の結果を記憶部１２０に格納する。また、第２クラスタ処理部１５０は、記憶部１２０をワークスペースとして、各種データを記憶部１２０に一時格納し、それらデータを利用して第２のクラスタ処理を実行する。

【００３６】

決定部１６０は、記憶部１２０に格納されている第１クラスタ処理部１４０および第２クラスタ処理１５０によるそれぞれの処理結果を基に、処理対象情報のクラスタ処理結果を決定する。本実施形態において、決定部１６０は、第２クラスタ処理部１５０によるクラスタ処理結果のうち第１クラスタ処理部１４０によるクラスタ処理結果からクラスタが変更されたものを除外し、それによる第２クラスタ処理１５０によるクラスタ処理結果を最終クラスタ処理結果と決定する。なお、決定部１６０は、第１クラスタ処理部１４０による処理結果のうち第２クラスタ処理部１５０によるクラスタ処理結果からクラスタが変更されたものを除外し、それによる第１クラスタ処理部１４０によるクラスタ処理結果を最終クラスタ処理結果と決定してもよい。つまり、決定部１６０は、第１クラスタ処理部１４０の処理結果と第２クラスタ処理１５０による処理結果間でクラスタが一致する情報を最終クラスタ処理結果と決定すればよい。決定部１６０は、最終クラスタ処理結果を記憶部１２０に格納する。

【００３７】

出力部１７０は、記憶部１２０に格納されている最終クラスタ処理結果を出力する。出力部１７０は、例えば、図２における入力デバイス４２および出力デバイス５２を使ってコンピュータ１００を使用しているユーザからの要求に応じて、最終クラスタ処理結果を出力デバイス５２に出力してもよい。また、出力部１７０は、例えば、図２における通信インターフェース６０を介して外部から受信した要求に応じて、要求元の装置に最終クラスタ処理結果を通信インターフェース６０を介して出力してもよい。

【００３８】

次に、前述した図３に示す各機能ブロックによる処理フローを説明する。図４は、処理全体のフローチャートを示す。なお、各機能ブロックは、制御部１１０により所定のタイミングで動作有効に設定されるものとする。この所定のタイミングは、本処理の前、処理を開始すべき時点などである。また、本説明においては、制御部１１０による制御内容を割愛または各機能ブロックによる処理内容の一部として説明している。制御部１１０と他

の各機能ブロックのそれぞれの処理機能は、実施の形態に応じて設定・配分されればよい。

【００３９】

まず、取得部１３０は、処理対象情報を取得する（Ｓ１００）。取得部１３０は、取得した処理対象情報を記憶部１２０に格納する。

【００４０】

第１クラスタ処理部１４０は、第１のクラスタ処理の実行開始タイミングか否かを判定する（Ｓ２００）。第１クラスタ処理部１４０は、実行開始タイミングであると判定すると（Ｓ２００；Ｙｅｓ）、処理対象情報を対象に第１のクラスタ処理を開始する。一方、実行開始タイミングでないと判定すると（Ｓ２００；Ｎｏ）、第１クラスタ処理部１４０は次の判定タイミングまで判定処理を待つ。例えば、第１クラスタ処理部１４０は、記憶部１２０に格納された処理対象情報が第１のクラスタ処理を行うために必要な情報量であるか否かを判定することによって、第１のクラスタ処理の実行開始タイミングか否かを判定してもよい。

【００４１】

第１クラスタ処理部１４０は、実行開始タイミングであると判定した場合、記憶部１２０に格納されている処理対象情報を対象に第１のクラスタ処理を実行する（Ｓ３００）。第１クラスタ処理部１４０は、第１のクラスタ処理の結果を記憶部１２０に格納する。

【００４２】

第２クラスタ処理部１５０は、第２のクラスタ処理の実行開始タイミングか否かを判定する（Ｓ４００）。第２クラスタ処理部１５０は、実行開始タイミングであると判定すると（Ｓ４００；Ｙｅｓ）、記憶部１２０に格納されている処理対象情報を対象に第２のクラスタ処理を開始する。一方、実行開始タイミングでないと判定すると（Ｓ４００；Ｎｏ）、第２クラスタ処理部１５０は次の判定タイミングまで判定処理を待つ。例えば、第２クラスタ処理部１５０は、第１のクラスタ処理が行われたか否かを判定することによって、第２のクラスタ処理の実行開始タイミングか否かを判定してもよい。

【００４３】

第２クラスタ処理部１５０は、実行開始タイミングであると判定した場合、記憶部１２０に格納されている第１のクラスタ処理の結果を使って、第２のクラスタ処理を実行する（Ｓ５００）。第２クラスタ処理部１５０は、第２のクラスタ処理の結果を記憶部１２０に格納する。第２クラスタ処理部１５０は、処理対象情報を対象にした第２のクラスタ処理を実行する場合、例えば、第１のクラスタ処理の結果を第２のクラスタ処理の入力情報および初期設定情報として使用する。

【００４４】

決定部１６０は、記憶部１２０に格納された第１のクラスタ処理の結果および第２のクラスタ処理の結果に基づいて、処理対象情報に対する最終のクラスタ処理結果を決定する（Ｓ６００）。決定部１６０は、最終のクラスタ処理結果を記憶部１２０に格納する。例えば、決定部１６０は、第２のクラスタ処理の結果のうち第１のクラスタ処理の結果からクラスタが変更されたものを除外し、それによる第２のクラスタ処理の結果を最終クラスタ処理結果として決定してもよい。

【００４５】

出力部１７０は、決定部１６０により記憶部１２０に格納された最終のクラスタ処理結果を出力する（Ｓ７００）。出力部１７０は、最終のクラスタ処理結果の要求元に応じた形態にて最終のクラスタ処理結果を出力する。例えば、出力部１７０は、前述の通信インターフェース６０を介してコンピュータ５００から処理要求を受信していた場合、通信インターフェース６０を介してコンピュータ５００に最終のクラスタ処理結果を送信するよう処理を実行してもよい。また、例えば、コンピュータ１００の入力デバイス４２を使って処理要求が指示されていた場合、出力インターフェース５０を介して出力デバイス５２に最終のクラスタ処理結果を出力するよう処理を実行してもよい。この出力処理により、本処理は終了する。なお、本処理は、停止指示を受けるまでＳ１００～Ｓ７００間の処理

が繰り返し実行されてもよい。

【００４６】

以上、本処理の説明をしたが、次のような形態で処理が行われるようにしてもよい。

【００４７】

例えば、第１クラスタ処理部１４０は、実行開始タイミングか否かを判定する際に、第１クラスタ処理部１４０が記憶部１２０に格納される処理対象情報を監視しそのタイミング判定を行ってもよい。また、実行開始タイミングに到達した旨を取得部１３０が第１クラスタ処理部１４０に通知するように構成し、それによって、第１クラスタ処理部１４０が実行開始タイミングを判定するようにしてもよい。

【００４８】

また、例えば、第１クラスタ処理部１４０および第２クラスタ処理１５０それぞれは別プロセスとして並列動作するようにしてもよく、第１クラスタ処理部１４０および第２クラスタ処理１５０によるそれぞれの実行開始タイミングの判定が並列に行われるようにしてもよい。

【００４９】

次に、以上説明した各処理の一例を説明する。以下に説明する各処理は一例であり、本発明はこの処理内容に限定されるものではない。なお、以下説明においては、図２におけるサーバ３００を処理対象情報の提供元、コンピュータ５００をコンピュータ１００の管理装置としている。また、図５を適宜参照して、各処理に関する一例を説明する。

【００５０】

図６は、取得処理の一例のフローチャートを示す。図６は、図３における取得部１３０により実施される処理であって、処理対象情報の取得方法の一例を示す。

【００５１】

まず、取得処理の開始が指示されると、取得部１３０は、コンピュータ５００により設定された情報に基づいて、サーバ３００にアクセスする（Ｓ１１０）。コンピュータ５００により設定される情報は、例えば、処理対象情報の提供元の識別情報やアドレス（本例ではサーバ３００を示す）、提供元へのアクセス形態などである。

【００５２】

取得部１３０は、サーバ３００と正常に接続されると、サーバ３００に処理対象情報となる情報を要求する（Ｓ１２０）。

【００５３】

その要求に応答してサーバ３００から情報が送信されると、取得部１３０は、その情報を受信し（Ｓ１３０）、記憶部１２０に格納する（Ｓ１４０）。このようにして、処理対象情報が取得される。

【００５４】

なお、この情報の取得処理は、停止指示があるまで繰り返し実行されるものであってもよい。例えば、サーバ３００が、ソーシャルメディアの一例であり投稿サービスであるツイッターのサービスを提供している場合、Ｓｔｒｅａｍｉｎｇ　ＡＰＩ（Ａｐｐｌｉｃａｔｉｏｎ　Ｐｒｏｇｒａｍ　Ｉｎｔｅｒｆａｃｅ）を利用すればそれが実現できる。このＳｔｒｅａｍｉｎｇ　ＡＰＩを利用してサーバ３００にアクセスすると、Ｓ１２０による情報の要求以後、サーバ３００は、継続して最新の投稿情報（ツイート（登録商標））を要求元の装置（本例ではコンピュータ１００）に順次送信する。したがって、取得部１３０は、継続して最新の投稿情報を順次受信することになる。取得部１３０は、このようにして順次受信した情報を記憶部１２０に順次格納すればよい。また、個々の情報は、他の情報と区別可能な識別情報や区切情報を含む。また、個々の情報がツイート（登録商標）である場合、その情報は固有のＵＲＬ（Ｕｎｉｆｏｒｍ　Ｒｅｓｏｕｒｃｅ　Ｌｏｃａｔｏｒ）で示される。従って、各ツイート（登録商標）のＵＲＬは、各情報を区別する識別情報として利用可能である。他の処理（例えば、第１のクラスタ処理部１４０による第１のクラスタ処理）では、このような情報を使って、情報の単位を認識できる。なお、取得部１３０が情報の単位を認識できるように、取得した情報にシリアル番号等の固有の識別

情報を付与して管理してもいいし、記憶部１２０における各情報の記憶アドレスを管理したテーブルを作成して情報単位を把握できるようにしてもよい。

【００５５】

図５の（Ａ）は、前述のようにして取得された処理対象情報の一例を示す。処理対象情報として、「Ａ社ＰＣを購入」、「Ａ社のＰＣ」、「Ａ社の株価」、「Ａ社株価上昇」の各テキスト情報が取得された例である。

【００５６】

図７は、第１クラスタ処理の一例のフローチャートを示す。図７は、図３における第１クラスタ処理部１４０により実施される処理であって、処理対象情報に対する第１のクラスタ処理の一例を示す。なお、本例では、前述のように、第１のクラスタ処理のクラスタ手法をキーワード抽出処理としている。

【００５７】

前述のように、第１クラスタ処理部１４０は、実行開始タイミングであるか否かを判定し、実行開始タイミングである場合に本処理の実行を開始する。なお、処理対象情報は複数個の情報を含むものとする。この実行開始タイミングは、例えば、前述のように、第１クラスタ処理部１４０による第１のクラスタ処理を行うために必要な情報量を取得したことをそのタイミングとすればよい。例えば、ツイート（登録商標）を処理対象情報とする場合は、その必要な情報量を数万件～数十万件に設定すればよい。この必要な情報量は、処理対象情報の内容や取得頻度などを基に適宜変更すればよい。

【００５８】

まず、第１クラスタ処理部１４０は、処理対象情報の中からキーワードを抽出する（Ｓ３１００）。次に、第１クラスタ処理部１４０は、抽出したキーワード別に、処理対象情報の各情報による集合を作成する（Ｓ３２００）。そして、第１クラスタ処理部１４０は、キーワード数を基に処理対象情報の各情報のベクトル情報を作成する（Ｓ３３００）。このベクトル情報は、処理対象情報の各情報の特徴を示す。

【００５９】

図８は、第１クラスタ処理におけるキーワード抽出処理の一例のフローチャートを示す。この図８に示されるキーワード抽出処理は、処理対象情報の中のすべてのキーワードの抽出処理、およびキーワードとクラスタ番号の対応情報の作成処理を含む。

【００６０】

まず、第１クラスタ処理部１４０は、記憶部１２０から処理対象情報のうちの１つの情報を取得する（Ｓ３１１０）。そして、第１クラスタ処理部１４０は、取得した情報について形態素解析を行う（Ｓ３１２０）。

【００６１】

第１クラスタ処理部１４０は、形態素解析により得られた情報の形態素の中から名詞を判別し、名詞をキーワードの候補として抽出する（Ｓ３１３０）。なお、記憶部１２０には、確定されたキーワードがクラスタ番号と対応付けて登録される。例えば、キーワードとクラスタ番号の対応情報は、記憶部１２０において、テーブル形式や配列形式など種々の形態で記憶されうる。また、本例においては、クラスタ番号は、最小番号を１とした正の整数値である。

【００６２】

第１クラスタ処理部１４０は、記憶部１２０に登録されている対応情報を参照して、キーワードの候補のうち記憶部１２０にキーワードとして登録されていない候補を特定する（Ｓ３１４０）。第１クラスタ処理部１４０は、特定した候補を新たなキーワードとし、クラスタ番号を更新しつつ（最終のクラスタ番号をインクリメントしつつ）、キーワードとクラスタ番号の新たな対応情報を記憶部１２０に登録する（Ｓ３１５０）。

【００６３】

第１クラスタ処理部１４０は、処理対象情報の各情報を対象にして、以上説明したＳ３１１０～Ｓ３１５０の処理を実行する。その後、第１クラスタ処理部１４０は、前述の処理によって抽出したキーワードの個数を記憶部１２０に登録する（Ｓ３１６０）。以後、

この記憶部１２０に登録されたキーワードの個数をキーワードの個数情報と称する。例えば、第１クラスタ処理部１４０は、キーワードの個数情報として最終のクラスタ番号を記憶部１２０に登録すればよい。

【００６４】

以上の処理によって、処理対象情報の中のすべてのキーワードの抽出、およびキーワードとクラスタ番号の対応情報の作成が完了する。

【００６５】

図９は、第１クラスタ処理における集合作成処理の一例のフローチャートを示す。この図９に示される集合作成処理は、キーワードごと（クラスタ番号ごと）に、キーワードを含む情報の集合を作成する処理を含む。

【００６６】

まず、第１クラスタ処理部１４０は、記憶部１２０に登録された対応情報の中から１つの対応情報を取得する（Ｓ３２１０）。なお、この取得処理が初回であれば、第１クラスタ処理部１４０は、クラスタ番号が１の対応情報を記憶部１２０から取得する。第１クラスタ処理部１４０は、取得した対応情報に含まれるキーワードが処理対象情報の各情報に含まれるか否かを判定する（Ｓ３２２０）。第１クラスタ処理部１４０は、処理対象情報の中でキーワードを含んでいる全ての情報をそのキーワードに対応する集合とする（Ｓ３２３０）。第１クラスタ処理部１４０は、作成された集合とクラスタ番号を対応付けて記憶部１２０に登録する。

【００６７】

第１クラスタ処理部１４０は、記憶部１２０に登録されているすべての対応情報を対象にして、以上説明したＳ３２１０～Ｓ３２３０の処理を実行する。これによって、キーワード別に情報がまとめ上げられ、情報の集合が作成される。

【００６８】

図５の（Ｂ）は、第１クラスタ処理部１４０による前述の処理結果の一例を示す。前述の処理によって記憶部１２０に登録された情報は、キーワード、テキスト集合、クラスタ番号を含む。なお、図５においては、前述の集合に対応するテキスト情報群をテキスト集合と称している。本図は、図８に示すキーワード抽出処理によって、処理対象情報から、「Ａ社」、「ＰＣ」、「株価」の各キーワードが抽出された例を示す。また、本図は、そのキーワード抽出処理によって、キーワード「Ａ社」にクラスタ番号「１」、キーワード「ＰＣ」にクラスタ番号「２」、キーワード「株価」にクラスタ番号「３」が対応付けられたことを示す。また、本図は、図９に示す集合作成処理によって、各キーワードに該キーワードを含むテキスト情報が対応付けて登録されたことを示す。図５の（Ｂ）においては、第１クラスタ処理部１４０の処理によって８つのエントリが記憶部１２０に登録されていることを示す。

【００６９】

図１０は、第１クラスタ処理におけるベクトル情報作成処理の一例のフローチャートを示す。この図１０に示されるベクトル情報作成処理は、処理対象情報の各情報についてベクトル情報を作成する処理を含む。

【００７０】

第１クラスタ処理部１４０は、キーワードの個数を特定する（Ｓ３３１０）。第１クラスタ処理部１４０は、例えば、図８のキーワード抽出処理において記憶部１２０に登録されたキーワードの個数情報を取得すればよい。なお、記憶部１２０にキーワードの個数情報が登録されていない場合、第１クラスタ処理部１４０は、図８のキーワード抽出処理によって作成され記憶部１２０に登録されている対応情報に含まれるクラスタ番号の中から最大のクラスタ番号を取得することによってキーワードの個数を特定できる。

【００７１】

第１クラスタ処理部１４０は、各キーワードを要素とした処理対象情報の多次元配列を記憶部１２０に作成する（Ｓ３３２０）。例えば、処理対象情報の情報数をｉ、ｎ（ｉ）をi番目の処理対象の情報に含まれるキーワード数とすると、その配列x_iは$x_i = (x_i$

$_{,1}$，$x_{i,2}$，・・・，$x_{i,n(i)}$）と表現できる。なお、$x_{i,j}$（１＜＝ｊ＜＝ｎ（ｉ））は、各キーワードに対応する。クラスタ番号を示し、クラスタ番号対応のキーワードを意味する。第１クラスタ処理部１４０は、処理対象情報の全ての情報について処理した後、各キーワードが出現する情報数を計算し、指定された閾値以上の出現回数のキーワードを選択し、選択されたキーワードの種類数をクラスタの数ｋとする。また、第１クラスタ処理部１４０は、同じキーワードを含む情報の集合を１つのクラスタとする。また、第１クラスタ処理部１４０は、選択された各キーワードに数字を付与しクラスタ番号とする。これらクラスタ数とクラスタ情報果は、第２クラスタ処理部１５０にて利用される。

【００７２】

第１クラスタ処理部１４０は、作成した配列x_iを使い、処理対象情報の各情報のベクトル情報を作成する（Ｓ３３３０）。その後、第１クラスタ処理部１４０は、この配列x_iの各要素であるキーワードが対応するベクトルの次元に対して値を埋める。ベクトルの各次元の値としては、例えば、２値情報（例えば、出現しない場合は値０、出現する場合は値１）、出現頻度、ＴＦ・ＩＤＦのような重みづけ手法などを用いればよい。この処理を処理対象情報の各情報について実施することで、各情報に対応するベクトル情報が作成される（x_iのベクトルが完成する）。第１クラスタ処理部１４０は、作成したベクトル情報を記憶部１２０に登録する。

【００７３】

次に、第２クラスタ処理の一例を説明する。

【００７４】

図１１は、第２クラスタ処理の一例のフローチャートを示す。図１１は、図３における第２クラスタ処理部１５０により実施される処理であって、第１のクラスタ処理の処理結果を用いた第２のクラスタ処理の一例を示す。なお、本例では、前述のように、第２のクラスタ処理のクラスタ手法をＫ－ｍｅａｎｓとしている。

【００７５】

前述のように、第２クラスタ処理部１５０は、第１のクラスタ処理が行われたか否かを判定し、実行開始タイミングである場合に本処理の実行を開始する。

【００７６】

まず、第２クラスタ処理部１５０は、記憶部１２０に格納されているキーワードの個数情報および各情報のベクトル情報を取得する（Ｓ６１０）。このキーワードの個数情報は、図８に示すキーワード抽出処理において、第１クラスタ処理部１４０により作成され、記憶部１２０に登録されたものである。また、各情報のベクトル情報は、図１０に示すベクトル情報作成処理において、第１クラスタ処理部１４０により作成され、記憶部１２０に登録されたものである。

【００７７】

第２クラスタ処理部１５０は、Ｓ６１０にて取得した情報を用いて第２クラスタ処理の一例であるＫ－ｍｅａｎｓのクラスタ処理における初期設定を行う（Ｓ６２０）。

【００７８】

ここで、一般的なＫ－ｍｅａｎｓによるクラスタ処理を説明する。

【００７９】

例えば、Ｋ－ｍｅａｎｓは以下の式１で示される目的関数を最小化する分割最適化クラスタ処理である。

式１：

$$\mathrm{Err}(\{X_i\}) = \sum_{i}^{k} \sum_{x \in X_i} \left\| x - \overline{x_i} \right\|^2$$

【００８０】

ここで、Ｘは、データ集合であり、ベクトルで表現されたデータｘの集合である。

$\{X_i\}$

はｋ個のクラスタ、

X_i

はi番目のクラスタであり、データ集合の網羅的で互いに疎な部分集合である。また、ｋはクラスタ数であり、

x_i

はセントロイド（重心）である。

【００８１】

上記を実現するために、一般的なＫ－ｍｅａｎｓは、以下に示す処理を行う。

【００８２】

まず、初期設定では、データ集合をランダムにｋ個のクラスタに分割し、それを初期クラスタとする。

【００８３】

その後、各クラスタについて以下の式２で示されるセントロイドの計算を行う。

式２：

$$\overline{x_i} = \frac{1}{|x_i|} \sum_{x \in X_i} x$$

【００８４】

$|x_i|$

はクラスタ

X_i

に含まれるデータ数である。

【００８５】

続いて、各データ

$x \in X$

において、各クラスタのセントロイド

x_i
との距離
$x \in X$
を計算し、距離が最小であるクラスタ
X_i
を見つけ、データをそのクラスタに割り当てる。このようにして、全てのデータがクラスタに割り当てると、式２においてセントロイドを更新する。

【００８６】

前述の各クラスタのセントロイド計算とクラスタへの割り当て処理は、反復数が設定回数に達するまで繰り返される。反復数が設定回数に達した場合、繰り返し処理が終了され、本アルゴリズムの最終処理結果としてｋ個のクラスタ
$\{X_i\}$
が出力される。なお、反復数が設定回数に達する前に、クラスタの割り当てが前回の反復時のクラスタの割り当てから変化がなかった場合に、繰り返し処理が終了されるようにしてもよい。このアルゴリズムによる計算量は、データ数をＮ、反復回数を定数とすると、Ｏ(Ｎｋ)となる。

【００８７】

一般的に、Ｋ－ｍｅａｎｓでは、初期クラスタを何回か変更して各初期クラスタにて前述のアルゴリズムを実行し、それぞれのアルゴリズム実行において前述の目的関数を最小化する分割を選択する。そして、この数回のアルゴリズムの実行に基づいて、大域最適に近い解の探索が行われる。

【００８８】

このように、Ｋ－ｍｅａｎｓを用いた一般的なクラスタ処理は、初期化においてデータ集合をランダムにクラスタ分割する。また、このようなデータ集合をランダムにクラスタ分割する初期クラスタ処理が何度か行われ（つまり、初期クラスタを何度か変更して）、前述のアルゴリズムが実行され解の探索が行われる。このため、計算量が多く、最終的な処理結果を得るために長い処理時間が必要とされる。

【００８９】

また、前述のように計算量はデータ数に比例することから、例えば処理対象情報が大規模データであると、非常に長い処理時間が必要とされる。そのため、処理対象情報を取得し、リアルタイムにクラスタ情報を提供するというような処理が困難となる。

【００９０】

本実施形態では、初期設定に第１クラスタ処理部１４０の処理結果を用いる。つまり、第２クラスタ処理部１５０は、前述の目的関数における初期クラスタのクラスタ数ｋに、前述で取得されたキーワードの個数情報を設定し、各クラスタにキーワード対応のクラスタ番号を付与する。また、第２クラスタ処理部１５０は、前述の目的関数におけるデータ集合Ｘに、前述で取得された各情報のベクトル情報を設定する。また、本実施形態では、第1クラスタ処理部１４０で決定したクラスタ情報を基にＫ－ｍｅａｎｓのセントロイドの初期値の計算を行う。このセントロイドの初期値の計算は、通常のＫ－ｍｅａｎｓと同

様に前述の式２により行われるが、一般的なＫ－ｍｅａｎｓでの処理のように初期のセントロイド計算のためのクラスタ作成がランダムで実施されるのではなく、本実施形態では、第１クラスタ処理部１４０で決定されたクラスタ情報を基に計算を行う。これにより、第1クラスタ処理部での処理結果が第２クラスタ処理部１５０に引き継がれる。

【００９１】

このように初期設定が完了した後、第２クラスタ処理部１５０は、Ｋ－ｍｅａｎｓによるクラスタ処理を実行する（Ｓ６３０）。この第２クラスタ処理部１５０におけるＫ－ｍｅａｎｓによるクラスタ処理では、前述の繰り返し処理が実行され、反復数が設定回数に達した場合に繰り返し処理を終了する。なお、前述のように、設定回数に達する前に、今回のクラスタの割り当てが前回の反復によるクラスタの割り当てから変化がなかった場合に、繰り返し処理が終了されるようにしてもよい。

【００９２】

前述のようにＫ－ｍｅａｎｓを用いた一般的なクラスタ処理では何度か初期クラスタを変更して前述のアルゴリズムが実行されるが、本実施形態においては、第２クラスタ処理１５０は前述のキーワードの個数情報を用いて決定した初期クラスタにて前述のアルゴリズムを１度だけ実行する。このため、本実施形態によれば、一般的なＫ－ｍｅａｎｓによるクラスタ処理と比較して、計算量が少なくなり処理時間の短縮が図れる。

【００９３】

第２クラスタ処理部１５０は、Ｋ－ｍｅａｎｓによる前述のクラスタ処理の結果を記憶部１２０に登録する（Ｓ６４０）。このクラスタ処理の結果は前述のキーワードに対応付けられた各情報に対するクラスタ番号であり、第２クラスタ処理部１５０は、この処理結果であるクラスタ番号を各情報に対応づけて記憶部１２０に登録する。

【００９４】

図５の（Ｃ）は、第２クラスタ処理部１５０による前述の処理結果の一例を示す。本図において、テキスト集合、クラスタ番号の情報は、前述の第１クラスタ処理部１４０による処理結果を示す。本図における新クラスタ番号は、第２クラスタ処理部１５０による前述の処理によって得られ、テキスト集合の各テキスト情報に対応付けて登録されたクラスタ番号である。

【００９５】

図１２は、決定処理の一例のフローチャートを示す。図１２は、図３における決定部１６０により実施される処理であって、第１クラスタ処理部１４０と第２クラスタ処理部１５０のそれぞれの処理結果に基づき、処理対象に対する最終のクラスタ処理結果の決定方法の一例を示す。

【００９６】

まず、第２クラスタ処理部１５０による処理が終了すると、決定部１６０は、前述のようにして登録された第１クラスタ処理部１４０と第２クラスタ処理部１５０の各処理結果のうち１つの対の情報を記憶部１２０から取得する（Ｓ７１０）。この対の情報は、第１クラスタ処理部１４０の処理によって得られたクラスタ番号と第２クラスタ処理部１５０の処理によって得られたクラスタ番号である。

【００９７】

そして、決定部１６０は、取得した２つのクラスタ番号を比較し、第２クラスタ処理部１５０のクラスタ処理により得られたクラスタ番号が第１クラスタ処理部１４０の処理により得られたクラスタ番号から変更されているか判定する（Ｓ７２０）。クラスタ番号が変更されている場合、決定部１６０は、それらクラスタ番号と、それらに対応する情報を記憶部１２０から削除する。

【００９８】

以上の処理を図５の（Ｃ）を用いて以下に説明する。

【００９９】

決定部１６０は、図５の（Ｃ）に示す記憶部１２０の登録情報のなかから、テキスト情報に対応するクラスタ番号（第１クラスタ処理部１４０の処理結果）と新クラスタ番号（

第２クラスタ処理部１５０）を取得する。例えば、図５の（Ｃ）において最初のエントリであるテキスト情報「Ａ社のＰＣを購入」を例にすると、決定部１６０は、そのテキスト情報に対応するクラスタ番号「１」と新クラスタ番号「２」を取得する。この取得された２つのクラスタ番号は異なるため、決定部１６０は、テキスト情報「Ａ社のＰＣを購入」、このテキスト情報に対応するクラスタ番号「１」および新クラスタ番号「２」を記憶部１２０から削除する。

【０１００】

決定部１６０は、記憶部１２０に登録された第１クラスタ処理部１４０と第２クラスタ処理部１５０の各対の情報（図５の（Ｃ）における各エントリ）を対象にして、以上説明したＳ７１０～Ｓ７３０の処理を実行する。

【０１０１】

すべての情報を対象に処理を実行した後、決定部１６０は、削除されず記憶部１２０に残っている各情報（各エントリ）を処理結果として記憶部１２０の処理結果格納領域に登録する（Ｓ７４０）。なお、すべての情報を対象にした処理の実行後ではなく、決定部１６０は、クラスタ番号の比較においてクラスタ番号が一致すると判定された際に、クラスタ番号の対およびそれに対応する情報を記憶部１２０の処理結果格納領域に登録するようにしてもよい。

【０１０２】

前述のように、決定部１６０は、第１クラスタ処理部１４０による処理結果と第２クラスタ処理部１５０による処理結果との間で、クラスタ番号が不一致となった情報を削除する。このクラスタ番号が不一致となった情報は、処理対象情報についてクラスタを決定する上で不安定な情報といえる。例えば、大規模データは多様な情報内容を含む。このような多様な情報内容のデータを処理対象情報としてクラスタ処理を行う場合、クラスタを決定する上で不安定な情報が非常に多く出現することになる。

【０１０３】

本実施形態は、クラスタを決定する上で不安定な情報を特定し、先行技術のように不安定な情報を詳細分類するのではなく、不安定な情報を削除してクラスタ処理の対象外とする。このような処理によって、本実施形態は高速なクラスタ処理を実現する。本実施形態は、多様な情報内容を含む例えば大規模データを対象にしたクラスタ処理に特に有効である。

【０１０４】

図５の（Ｃ）では、８つのエントリのうち上部の４つのエントリにおいて、第１クラスタ処理部１４０の処理によって得られたクラスタ番号と、第２クラスタ処理部１５０の処理によって得られた新クラスタ番号とが相違する。したがって、図５の（Ｃ）に示す例の場合、この４つのエントリが決定部１６０によって削除される。図５の（Ｄ）は、決定部１６０の処理によって、クラスタ番号が不一致のエントリが削除された結果を示す。言い換えれば、図５の（Ｄ）は、決定部１６０の処理によって、記憶部１２０の処理結果格納領域に登録されたエントリを示す。

【０１０５】

決定部１６０によって記憶部１２０の処理結果格納領域に登録された最終クラスタ処理結果は、出力部１７０によって出力される。出力部１７０は、記憶部１２０の処理結果格納領域に登録されている最終クラスタ処理結果を、例えば、そのまま出力してもいいし、クラスタを区別可能にして各情報を出力してもよい。出力部１７０による出力形態は、要求元の装置や出力先の装置、処理結果の使用形態に応じて適宜変更すればよい。

【０１０６】

図５の（Ｅ）は、図５の（Ａ）に示す処理対象情報について本実施形態によるクラスタ処理が実行された結果を示す。図５の（Ｅ）は、図５の（Ｄ）の情報内容に基づくものである。図５の（Ｄ）に示されるように、処理対象情報の各テキスト情報である、「Ａ社ＰＣを購入」、「Ａ社のＰＣ」、「Ａ社の株価」、「Ａ社株価上昇」は、「Ａ社ＰＣを購入」と「Ａ社のＰＣ」に同じクラスタ番号「２」、「Ａ社の株価」と「Ａ社株価上昇」に同

じクラスタ番号「３」が付与されている。つまり、図５の（Ｅ）に示すように、「Ａ社ＰＣを購入」と「Ａ社のＰＣ」が１つのクラスタ、「Ａ社の株価」と「Ａ社株価上昇」が他のクラスタとなる。

【０１０７】

以上説明したように、本実施形態では、Ｋ－ｍｅａｎｓを例にした第２のクラスタ処理よりも高速な（短時間で処理結果を得られる）キーワード抽出処理を例にした第１のクラスタ処理にてまず処理対象情報を分類し（クラスタ処理し）、その分類結果を用いて第２のクラスタ処理にて分類した（クラスタ処理した）後、双方の分類結果で相違する情報を除外した分類結果を処理対象情報に対する最終処理結果と決定する。つまり、本実施形態は、処理スピード重視で第１のクラスタ処理にて大雑把な分類を行い、精度重視で第２のクラスタ処理にて分類した結果を用いて第１のクラスタ処理の結果の精度を評価し、精度の悪い情報を排除している。このようにすることで、本実施形態は、高速なクラスタ処理を実現しつつ、クラスタの精度を維持している。

【０１０８】

なお、前述した本実施形態では、第１のクラスタ処理をキーワード抽出処理、第２のクラスタ処理をＫ－ｍｅａｎｓとしたが、それぞれのクラスタ処理はこれに限定されるものではない。例えば、第２のクラスタ処理は混合正規分布推定であってもよい。

【０１０９】

以上の実施形態に関し、更に以下の付記を開示する。

（付記１）

コンピュータが、

第２のクラスタ手法と異なりかつ該第２のクラスタ手法よりも高速な第１のクラスタ手法によって対象情報に対して第１の分類処理を実行し、

前記第１の分類処理の結果に基づいて前記第２のクラスタ手法によって第２の分類処理を実行し、

前記第１の分類処理の結果と前記第２の分類処理の結果とに基づいて前記対象情報の分類を決定する、

ことを特徴とするクラスタ処理方法。

（付記２）

前記対象情報の分類を決定することは、前記第１の分類処理の結果と前記第２の分類処理の結果間で分類の変更がなかった前記第１の分類処理の結果または前記第２の分類処理の結果に基づいて前記対象情報の分類を決定する、

ことを特徴とする付記１のクラスタ処理方法。

（付記３）

前記第１の分類処理の結果は、前記対象情報に関連した分類個数情報を含み、

前記第２の分類処理は、前記分類個数情報を前記第２のクラスタ手法における初期情報として実行される、

ことを特徴とする付記１または２のクラスタ処理方法。

（付記４）

前記第１の分類処理の結果は、前記対象情報に含まれるキーワードを基にした前記対象情報の特徴情報を含み、

前記第２の分類処理は、前記対象情報の特徴情報を前記第２のクラスタ手法における前記初期情報として実行される、

ことを特徴とする付記３のクラスタ処理方法。

（付記５）

前記第１の分類処理は、前記対象情報を用いた前記キーワード別の集合を作成し、異なる前記キーワードに対応する前記集合を異なる分類とする、

ことを特徴とする付記４のクラスタ処理方法。

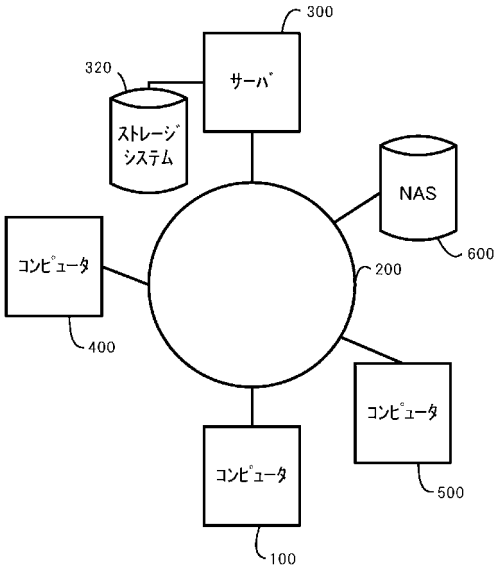
（付記６）

前記第２のクラスタ手法は、処理対象情報の類似性を基に処理対象情報を分類する手法

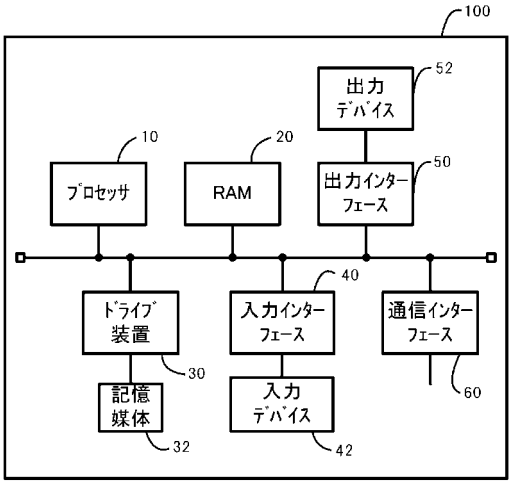
である、
ことを特徴とする付記１乃至５のいずれか１つのクラスタ処理方法。
（付記７）
第２のクラスタ手法と異なりかつ該第２のクラスタ手法よりも高速な第１のクラスタ手法によって対象情報に対して第１の分類処理を実行する第１処理部と、
前記第１の分類処理の結果に基づいて前記第２のクラスタ手法によって第２の分類処理を実行する第２処理部と、
前記第１の分類処理の結果と前記第２の分類処理の結果とに基づいて前記対象情報の分類を決定する決定部と、
を有することを特徴とするクラスタ処理装置。
（付記８）
前記決定部は、前記第１の分類処理の結果と前記第２の分類処理の結果間で分類の変更がなかった前記第１の分類処理の結果または前記第２の分類処理の結果に基づいて前記対象情報の分類を決定する、
ことを特徴とする付記７のクラスタ処理装置。
（付記９）
前記第１処理部による前記第１の分類処理の結果は、前記対象情報に関連した分類個数情報を含み、
前記第２処理部は、前記分類個数情報を前記第２のクラスタ手法における初期情報として前記第２の分類処理を実行する、
ことを特徴とする付記７または８のクラスタ処理装置。
（付記１０）
前記第１処理部による前記第１の分類処理の結果は、前記対象情報に含まれるキーワードを基にした前記対象情報の特徴情報を含み、
前記第２処理部は、前記対象情報の特徴情報を前記第２のクラスタ手法における前記初期情報として前記第２の分類処理を実行する、
ことを特徴とする付記９のクラスタ処理方法。
（付記１１）
前記第１処理部は、前記対象情報を用いた前記キーワード別の集合を作成し、異なる前記キーワードに対応する前記集合を異なる分類とする、
ことを特徴とする付記１０のクラスタ処理装置。
（付記１２）
前記第２のクラスタ手法は、処理対象情報の類似性を基に処理対象情報を分類する手法である、
ことを特徴とする付記７乃至１１のいずれか１つのクラスタ処理装置。
（付記１３）
コンピュータに、
第２のクラスタ手法と異なりかつ該第２のクラスタ手法よりも高速な第１のクラスタ手法によって対象情報に対して第１の分類処理を実行し、
前記第１の分類処理の結果に基づいて前記第２のクラスタ手法によって第２の分類処理を実行し、
前記第１の分類処理の結果と前記第２の分類処理の結果とに基づいて前記対象情報の分類を決定する、
処理を実行させるためのプログラム。
（付記１４）
前記対象情報の分類を決定することは、前記第１の分類処理の結果と前記第２の分類処理の結果間で分類の変更がなかった前記第１の分類処理の結果または前記第２の分類処理の結果に基づいて前記対象情報の分類を決定する、
ことを特徴とする付記１３のプログラム。
（付記１５）

前記第１の分類処理の結果は、前記対象情報に関連した分類個数情報を含み、
前記第２の分類処理は、前記分類個数情報を前記第２のクラスタ手法における初期情報として実行される、
ことを特徴とする付記１３または１４のプログラム。
（付記１６）
前記第１の分類処理の結果は、前記対象情報に含まれるキーワードを基にした前記対象情報の特徴情報を含み、
前記第２の分類処理は、前記対象情報の特徴情報を前記第２のクラスタ手法における前記初期情報として実行される、
ことを特徴とする付記１５のプログラム。
（付記１７）
前記第１の分類処理は、前記対象情報を用いた前記キーワード別の集合を作成し、異なる前記キーワードに対応する前記集合を異なる分類とする、
ことを特徴とする付記１６のプログラム。
（付記１８）
前記第２のクラスタ手法は、処理対象情報の類似性を基に処理対象情報を分類する手法である、
ことを特徴とする付記１３乃至１７のいずれか１つのプログラム。
【符号の説明】
【０１１０】
１：コンピュータシステム
１００、４００、５００：コンピュータ
２００：ネットワーク
３００：サーバ
３２０：ストレージシステム
６００：ＮＡＳ
１０：プロセッサ
２０：ＲＡＭ
３０：ドライブ装置
３２：記憶媒体
４０：入力インターフェース
４２：入力デバイス
５０：出力インターフェース
５２：出力デバイス
６０：通信インターフェース
７０：バス
１１０：制御部
１２０：記憶部
１３０：取得部
１４０：第１クラスタ処理部
１５０：第２クラスタ処理部
１６０：決定部
１７０：出力部

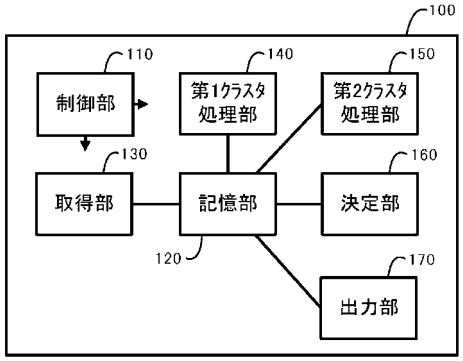
【図１】


300
サーバ
320
ストレージ
システム
NAS
600
コンピュータ
400
200
コンピュータ
500
コンピュータ
100

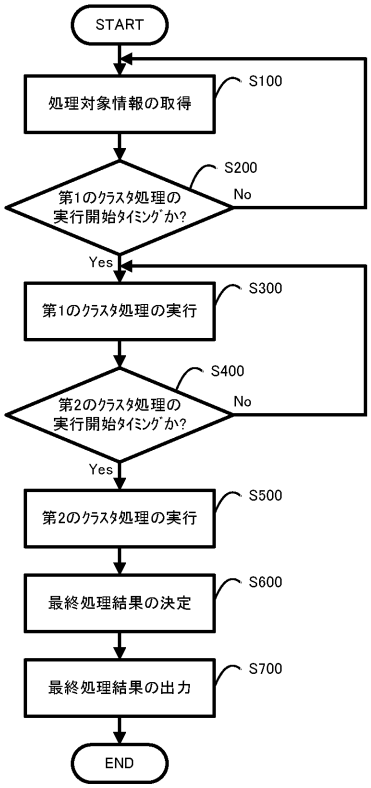
【図２】


100
出力
デバイス
52
10
20
プロセッサ
RAM
出力インター
フェース
50
ドライブ
装置
40
入力インター
フェース
通信インター
フェース
30
記憶
媒体
入力
デバイス
60
32
42

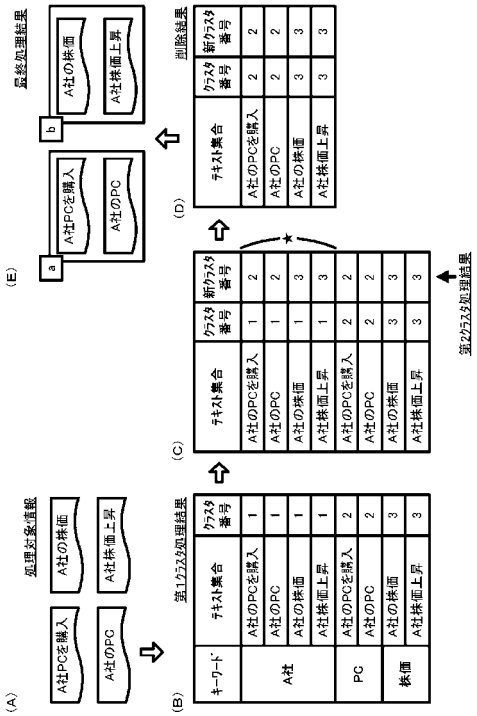
【図３】


100
110
140
150
制御部
第1クラスタ
処理部
第2クラスタ
処理部
130
160
取得部
記憶部
決定部
120
170
出力部

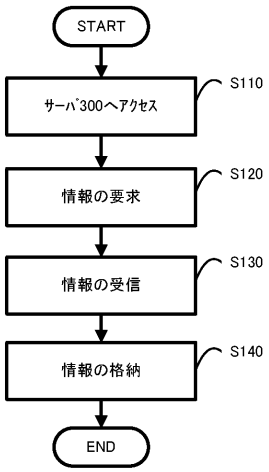
【図４】


START
処理対象情報の取得
S100
S200
第1のクラスタ処理の
実行開始タイミングか?
No
Yes
第1のクラスタ処理の実行
S300
S400
第2のクラスタ処理の
実行開始タイミングか?
No
Yes
第2のクラスタ処理の実行
S500
最終処理結果の決定
S600
最終処理結果の出力
S700
END

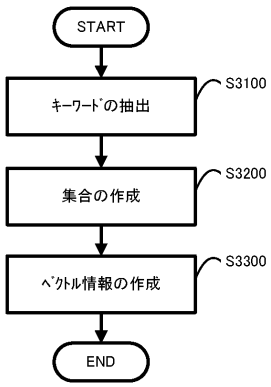
【図５】


(A)
処理対象情報
A社PCを購入
A社のPC
A社の株価
A社株価上昇
(B)
第1クラスタ処理結果
キーワード
テキスト集合
クラスタ番号
A社
PC
株価
A社のPCを購入
A社のPC
A社の株価
A社株価上昇
(C)
第2クラスタ処理結果
新クラスタ番号
(D)
削除結果
(E)
最終処理結果
a
b

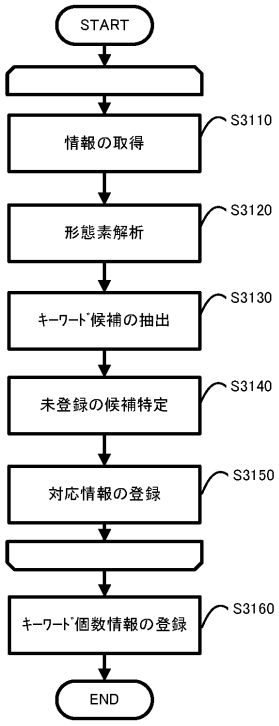
【図６】


START
サーバ300へアクセス
S110
情報の要求
S120
情報の受信
S130
情報の格納
S140
END

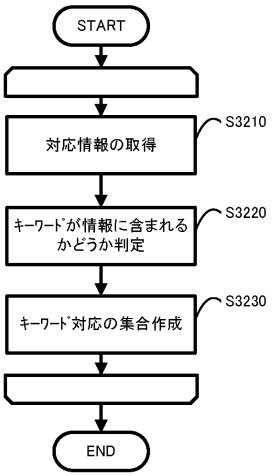
【図７】


START
キーワードの抽出
S3100
集合の作成
S3200
ベクトル情報の作成
S3300
END

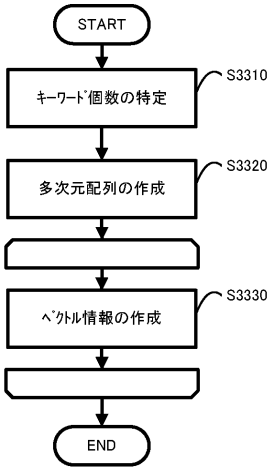
【図８】


START
情報の取得
S3110
形態素解析
S3120
キーワード候補の抽出
S3130
未登録の候補特定
S3140
対応情報の登録
S3150
キーワード個数情報の登録
S3160
END

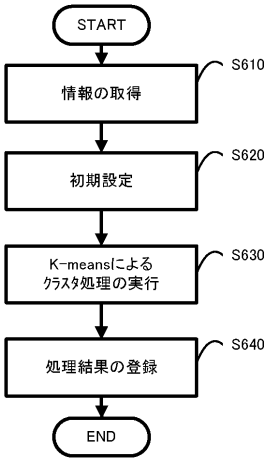
【図９】



【図１０】



【図１１】



【図１２】