

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2015年5月21日(21.05.2015)



(10) 国際公開番号

WO 2015/072085 A1

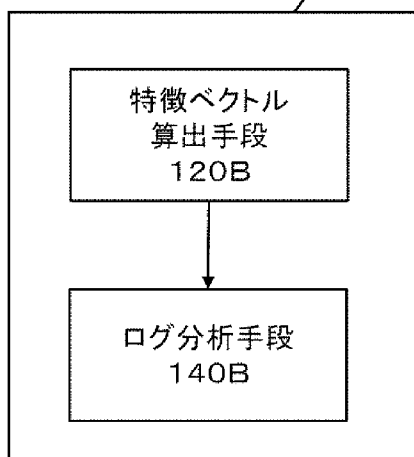
- (51) 国際特許分類:
G06F 11/34 (2006.01) G06F 11/30 (2006.01)
- (21) 国際出願番号: PCT/JP2014/005335
- (22) 国際出願日: 2014年10月21日(21.10.2014)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2013-233954 2013年11月12日(12.11.2013) JP
- (71) 出願人: 日本電気株式会社(NEC CORPORATION) [JP/JP]; 〒1088001 東京都港区芝五丁目7番1号 Tokyo (JP).
- (72) 発明者: 外川 遼介(TOGAWA, Ryosuke); 〒1088001 東京都港区芝五丁目7番1号日本電気株式会社内 Tokyo (JP).
- (74) 代理人: 下坂 直樹(SHIMOSAKA, Naoki); 〒1088001 東京都港区芝五丁目7番1号日本電気株式会社内 Tokyo (JP).
- (81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

[続葉有]

(54) Title: LOG ANALYSIS SYSTEM, LOG ANALYSIS METHOD, AND STORAGE MEDIUM

(54) 発明の名称: ログ分析システム、ログ分析方法、および、記憶媒体

ログ分析システム100B



(57) Abstract: An objective of the present invention is to appropriately classify log messages when analyzing the log messages. Provided is a log analysis system comprising: a feature vector computation means for computing a feature vector corresponding to a log message of interest among a set of log messages which is outputted from a system to be analyzed, according to whether the log message includes a prescribed text character string; and a log analysis means for analyzing the log message corresponding to the feature vector on the basis of the feature vector.

(57) 要約: ログメッセージを分析する際に、ログメッセージを適切に分類する。ログ分析システムは、分析対象システムから出力されたログメッセージの集合のうち注目するログメッセージに対応する特徴ベクトルを、当該ログメッセージが所定の文字列を含むか否かに応じて、算出する特徴ベクトル算出手段と、前記特徴ベクトルに基づいて、前記特徴ベクトルに対応する前記ログメッセージを分析するログ分析手段と、を備える。

- 100B Log analysis system
120B Feature vector computation means
140B Log analysis means

WO 2015/072085 A1



添付公開書類:

— 国際調査報告 (条約第 21 条(3))

明 細 書

発明の名称：

ログ分析システム、ログ分析方法、および、記憶媒体

技術分野

[0001] 本発明は、情報処理システムが出力するログを分析する技術に関する。

背景技術

[0002] コンピュータシステムの運用管理において、運用管理者は、障害などの異常の発生を監視し、その異常が発生した原因を分析する。そのために運用管理者は、当該コンピュータシステムが出力したログを分析する。しかし、大規模化かつ複雑化した現在のコンピュータシステムが出力するログのサイズ（または数量）は膨大であり、運用管理者はその全てを詳細に分析することはできない。

[0003] そこで、ログを分類するための分類ルールを定義し、コンピュータシステムにより出力されたメッセージを分類ルールに従って分類することが行われる。長年に亘って運用管理者が人手で分類ルールを定義していたが、膨大なサイズのログに対して網羅的な分類ルールを人手で定義することは困難である。そこで、近年においては、ログを分類するための分類ルールを自動的に作成する技術が提案されている。

[0004] 特許文献１は、分散システムが出力するメッセージが、当該分散システムが正常に運用されていることを示すメッセージであるのか、それとも当該分散システムに異常が発生したことを示すメッセージであるのかを、判定する技術の一例を開示する。以下、正常に運用されていることを示すメッセージを「正常なメッセージ」と記載する。また、異常が発生したことを示すメッセージを「異常なメッセージ」と記載する。

[0005] 特許文献１が開示する技術は、過去に分散システムが正常に運用されていた際に当該分散システムが出力したログを収集し、当該ログから特徴的なパターンを抽出する。特許文献１が開示する技術は、連続するメッセージの組

み合わせに着目して、当該特徴的なパターンを抽出する。特許文献１が開示する技術は、特徴的なパターンと、判定の対象となるメッセージと、を比較することにより、判定対象のメッセージが正常なメッセージであるのか否かを判定する。

[0006] 特許文献２は、通信を発生させた端末のアドレスに基づいて、当該通信が悪意のある通信か否かを判定するための技術の一例を開示する。特許文献２が開示する技術は、過去に悪意のある通信を発生させた端末のアドレスに特徴的な構造を、機械学習を適用して学習する。

[0007] 特許文献２が開示する技術は、判定対象である通信を発生させた端末のアドレスと、機械学習による学習結果と、を比較することにより、当該通信が悪意のある通信か否かを判定する。特許文献２はさらに、アドレスを構成するビット列の全部または一部を用いて、悪意のある通信を発生させる端末のアドレスに特徴的な構造を学習することを開示する。

先行技術文献

特許文献

[0008] 特許文献１：特開２００６－３１８０７１号公報

特許文献２：特開２０１２－１７５２９６号公報

発明の概要

発明が解決しようとする課題

[0009] 特許文献１が開示する技術は、連続するメッセージの組み合わせに着目して、分析対象のメッセージが正常であるのか異常であるのかを判定する。特許文献２が開示する技術は、アドレスを構成するビット列に着目して、分析対象のアドレスが正常であるのか異常であるのかを判定する。ログメッセージを分類する場合において、ログメッセージを構成する複数のログ要素のうち、どのログ要素に着目してログメッセージを分類するか、ということは、ログメッセージを適切に分類するために重要である。

[0010] ここで、ログ要素とは、例えば、ログメッセージを構成する、当該ログメ

ッセージが出力された時刻、ログID（識別子）、ログレベル、メッセージ本文、または、メッセージ本文の一部分などである。

[0011] ここで、以下の状況を仮定して、下記2つの具体例についてそれぞれ検討する。仮定する状況は、「正常なログメッセージと異常なログメッセージとの差異は、ログメッセージを構成する複数のログ要素のうち一部のログ要素に現れる」という状況である。

[0012] （具体例1）例えば、ログメッセージを構成する複数のログ要素のうち全てのログ要素に着目して、ログメッセージを分類することを考える。この場合、特許文献1が開示する技術は、正常なログメッセージと異常なログメッセージとの差異が現れないログ要素にまで着目してログメッセージを分類するため、ログメッセージを適切に分類することは難しい。

[0013] （具体例2）例えば、ログメッセージを構成する複数のログ要素のうち一部のログ要素に着目して、ログメッセージを分類することを考える。この場合、この場合、どのログ要素に着目してログメッセージを分類するかに応じて、ログメッセージが適切に分類されるか否かが変わってくる。不適切なログ要素に着目してログメッセージを分類してしまうと、ログメッセージを適切に分類することはできない。

[0014] 特許文献2は、アドレスを構成するビット列の一部を用いて悪意のある通信を発生させる端末のアドレスに特徴的な構造を学習する技術を開示する。しかし特許文献2は、アドレスを構成するビット列のどの部分に着目すべきかを選択する基準までは開示していない。

[0015] 上述した課題を解決するため、本発明は、ログメッセージを分析する際に、ログメッセージを適切に分類することを目的とする。

課題を解決するための手段

[0016] 本発明の第1の側面は、分析対象システムから出力されたログメッセージの集合のうち注目するログメッセージに対応する特徴ベクトルを、当該ログメッセージが所定の文字列を含むか否かに応じて算出する特徴ベクトル算出手段と、前記特徴ベクトルに基づいて、前記特徴ベクトルに対応する前記ロ

グメッセージを分析するログ分析手段と、を備え、前記特徴ベクトル算出部は、前記所定の文字列を、前記ログメッセージの集合に含まれる複数の文字列のうちから、前記ログメッセージの集合において当該文字列が出現する頻度に応じて選択する、ログ分析システムである。

[0017] 本発明の第2の側面は、コンピュータが、分析対象システムから出力されたログメッセージの集合のうち注目するログメッセージが、所定の文字列を含むか否かに応じて、前記注目するログメッセージに対応する特徴ベクトルを算出し、前記特徴ベクトルに基づいて、前記特徴ベクトルに対応する前記ログメッセージを分析し、前記ログメッセージの集合に含まれる複数の文字列のうちから、前記ログメッセージの集合において当該文字列が出現する頻度に応じて前記所定の文字列を選択する、ログ分析方法である。

[0018] 本発明の第3の側面は、コンピュータに、分析対象システムから出力されたログメッセージの集合のうち注目するログメッセージが、所定の文字列を含むか否かに応じて、前記注目するログメッセージに対応する特徴ベクトルを算出する処理と、前記特徴ベクトルに基づいて、前記特徴ベクトルに対応する前記ログメッセージを分析する処理と、前記ログメッセージの集合に含まれる複数の文字列のうちから、前記ログメッセージの集合において当該文字列が出現する頻度に応じて前記所定の文字列を選択する処理と、を実行させるプログラムである。

[0019] また、本発明の目的は、上記のプログラムが格納されたコンピュータ読み取り可能な記録媒体によっても達成される。

発明の効果

[0020] 本発明によれば、ログメッセージを分析する際に、ログメッセージを適切に分類することができる。

図面の簡単な説明

[0021] [図1]本発明の第1の実施形態における、ログ分析システム100の構成を示すブロック図である。

[図2]本発明の第1の実施形態における、ログファイル910の一例を示す図

である。

[図3]本発明の第1の実施形態における、特徴ベクトル算出部120の詳細を説明するブロック図である。

[図4]本発明の第1の実施形態における、文字列集計部122が集計した文字列の一例を示す図である。

[図5]本発明の第1の実施形態における、変換定義記憶部210が記憶する変換閾値情報の一例を示す図である。

[図6]本発明の第1の実施形態における、変換定義記憶部210が記憶する変換指標情報の一例を示す図である。

[図7]本発明の第1の実施形態における、ベクトル定義情報の一例を示す図である。

[図8]本発明の第1の実施形態における、特徴ベクトル算出部120が算出した特徴ベクトルの一例を示す図である。

[図9]本発明の第1の実施形態における、分類指標算出部130の動作を説明する図である。

[図10]本発明の第1の実施形態における、分類定義情報の一例を示す図である。

[図11]本発明の第1の実施形態における、ログ分析システム100が、ベクトル定義情報を決定する動作を説明するフローチャートである。

[図12]本発明の第1の実施形態における、ログ分析システム100が、分類定義情報を作成する動作を説明するフローチャートである。

[図13]本発明の第1の実施形態における、ログ分析システム100が、ログを分析する動作を説明するフローチャートである。

[図14]本発明のログ分析システム100のハードウェア構成の一例を説明するブロック図である。

[図15]本発明の第2の実施形態における、ログ分析システム100Aの構成を示すブロック図である。

[図16]本発明の第2の実施形態における、処理定義情報記憶部240が記憶

する情報の一例を示す図である。

[図17]本発明の第2の実施形態における、ログ分析システム100Aの動作を説明するフローチャートである。

[図18]本発明の第3の実施形態における、ログ分析システム100Bの構成を示すブロック図である。

発明を実施するための形態

[0022] 以下、本発明の実施形態を、図面を参照して詳細に説明する。

<第1の実施形態>

[構成の説明]

図1は、本発明の第1の実施形態にかかるログ分析システム100の構成を示すブロック図である。ログ分析システム100は、分析対象システム900が出力するログファイル910を分析する。

[0023] 図1に示すように、ログ分析システム100は、ログ収集部110と、特徴ベクトル算出部120と、分類指標算出部130と、ログ分析部140と、出力部150と、変換定義記憶部210と、ベクトル定義情報記憶部220と、分類定義情報記憶部230と、を備える。

[0024] ログ収集部110は、ログファイル910を収集する。ログ収集部110は、分析対象システム900からログファイル910を受信してもよい。または、ログ収集部110は、図示しない記憶部からログファイル910を読み出してよい。あるいは、ログ収集部110は、運用管理者からログファイル910の入力を受け付けてもよい。

[0025] ログ収集部110は、例えば、収集対象とするログファイル910の指定または日時の指定など、収集するログの範囲の指定を運用管理者から受け付けてもよい。あるいはログ収集部110は、ログメッセージの分析に必要な情報を定義した図示しないファイルを読み込み、ファイルが定義する情報に従って、取得したログファイル910の形式をログ分析システム100が分析しやすい形式に変換してもよい。

[0026] 特徴ベクトル算出部120は、ログファイル910と、変換定義記憶部2

10から読み出した情報と、に基づいて、ベクトル定義情報を決定する。ベクトル定義情報は、ログメッセージに基づいて当該ログメッセージの特徴を表す特徴ベクトルを算出する方法を定義する情報である。特徴ベクトル算出部120はベクトル定義情報をベクトル定義情報記憶部220に記憶する。

[0027] また、特徴ベクトル算出部120は、ベクトル定義情報に基づいて、ログファイル910を構成する複数のログメッセージのそれぞれに対して、当該ログメッセージに対応する特徴ベクトルを算出する。特徴ベクトル算出部120の詳細は後述する。

[0028] 分類指標算出部130は、特徴ベクトル算出部120が算出した特徴ベクトルに基づいて、ログメッセージを分類するための分類定義情報を作成する。分類指標算出部130は、分類定義情報を分類定義情報記憶部230に記憶する。分類指標算出部130の詳細は後述する。

[0029] ログ分析部140は、分類定義情報記憶部230から分類定義情報を読み出す。ログ分析部140は、分類定義情報と、特徴ベクトル算出部120が算出した特徴ベクトルとに基づいて、当該特徴ベクトルに対応するログメッセージを分析する。ログ分析部140は、ログメッセージを、分類定義情報が定義する複数の集合のうちいずれかの集合に分類する。あるいはログ分析部140は、当該ログメッセージを正常なログメッセージであるのか異常なログメッセージであるのか判定する。ログ分析部140の詳細は後述する。

[0030] 出力部150は、ログ分析部140が分析した結果を出力する。例えば、分析対象のログメッセージが既知の集合に属すると判定された場合、出力部150は、当該ログメッセージと当該集合とを関連付けて出力する。例えば、分類定義情報が定義するいずれの集合にも分析対象のログメッセージが属しないと判定された場合、出力部150は、当該ログメッセージを異常なログメッセージとして出力する。

[0031] 変換定義記憶部210は、変換閾値定義と、変換指標定義とを記憶する。変換閾値定義および変換指標定義は、特徴ベクトル算出部120がベクトル定義情報を決定する際の基準となる情報である。

- [0032] ベクトル定義情報記憶部 220 は、特徴ベクトル算出部 120 が決定したベクトル定義情報を記憶する。
- [0033] 分類定義情報記憶部 230 は、分類指標算出部 130 が作成した分類定義情報を記憶する。
- [0034] 分析対象システム 900 は、運用管理の対象である情報処理システムである。分析対象システム 900 は、分析対象システム 900 の利用状況、性能情報、または、データ通信の送受信状況などを記録したファイルであるログファイル 910 を出力する。分析対象システム 900 は、コンピュータシステムには限られず、例えば化学プラントまたは発電所等であってもよい。
- [0035] ログファイル 910 は、複数のログメッセージから構成される。言い換えると、ログファイルは複数のログメッセージの集合である。ログメッセージは、ログレコードとも呼ばれる。ログメッセージは、当該ログメッセージが出力された時刻、ログ ID、メッセージ本文、または、ログレベルなど、複数のログ要素により構成される。
- [0036] 図 2 は、ログファイル 910 の一例を説明する図である。ログメッセージは、図 2 に示した「行」に相当する情報である。例えば図 2 には、3 つのログメッセージが示されている。すなわち、
“130201 09:04:01 mysqld started”、
“130201 09:04:01 InnoDB : started;”、
“130201 09:04:01 [note] user/libexec/mysqld:”、
という 3 つのログメッセージが図 2 に示されている。
- [0037] 図 3 は、図 1 に示した特徴ベクトル算出部 120 の構成を更に詳細に説明したブロック図である。
- [0038] 図 3 に示すように、特徴ベクトル算出部 120 は、文字列抽出部 121 と、文字列集計部 122 と、基底ベクトル定義部 123 と、重み定義部 124 と、算出部 125 とを備える。
- [0039] <文字列抽出部 121>
文字列抽出部 121 は、ログファイル 910 を構成する複数のログメッセ

ージのそれぞれに対して、ログメッセージを構成する文字列を抽出する。

[0040] 上述したように、ログメッセージは、当該ログメッセージが出力された時刻、ログID、メッセージ本文、または、ログレベルなど、複数のログ要素により構成される。ログ分析システム100は、ログメッセージを構成するそれぞれのログ要素を、文字列または文字列の集合として取り扱う。

[0041] ここで、「文字列」という用語を定義する。本実施形態における「文字列」とは、ログメッセージを構成する文字の組み合わせのうち、意味を有する特定の複数文字による組み合わせを表す。本実施形態における「文字列」は、例えば「単語」である。本実施形態における「文字列」は、例えば「文法上、意味または職能をもった最小の言語単位」である。本実施形態における「文字列」は、フレーズまたは文章であってもよい。実施形態における「文字列」は、日付、時刻またはIP (Internet Protocol) アドレスなどを表す文字列であってもよい。実施形態における「文字列」は、「単語と、カッコなどの記号との、組み合わせ」であってもよい。「文字列」という用語の定義については、他の実施形態においても同様である。

[0042] ログメッセージを構成するある文字列と他の文字列との間は、例えば、カンマ、スペース、タブ、括弧、コロン、カンマ、ドットまたはスラッシュなどの記号で区切られていることを想定する。文字列抽出部121は、各ログメッセージを構成する文字列を、これらの記号で区切るにより抽出してもよい。文字列抽出部121は、抽出した文字列を正規表現により表現してもよい。

[0043] 文字列抽出部121の詳細を、具体例を用いて説明する。例えば、図2に示した3つのログメッセージから文字列を抽出すると、下記のような文字列が抽出される。なお、ここで示す具体例はあくまで発明の理解を容易にするためのものであり、文字列抽出部121の動作を限定して解釈するためのものではない。

[0044] (抽出された文字列の具体例) 130201, 09:04:01, mysqld, started, 130201, 09:04:01, InnoDB, started, 130201, 09:04:01, [note], user, libexe

c, mysqld.

[0045] 文字列抽出部 1 2 1 は、ログメッセージを構成する全てのログ要素から文字列を抽出する必要は必ずしも無い。文字列抽出部 1 2 1 は例えば、ログメッセージを構成するメッセージ本文のみから文字列を抽出してもよい。

[0046] <文字列集計部 1 2 2、基底ベクトル定義部 1 2 3、重み定義部 1 2 4>
文字列集計部 1 2 2、基底ベクトル定義部 1 2 3、および、重み定義部 1 2 4 は、文字列抽出部が抽出した文字列、並びに、変換定義記憶部 2 1 0 から読み込んだ変換閾値定義および変換指標定義に基づいて、ベクトル定義情報を決定する。重み定義部 1 2 4 は、ベクトル定義情報をベクトル定義情報記憶部 2 2 0 に記憶する。文字列集計部 1 2 2 と、基底ベクトル定義部 1 2 3 と、重み定義部 1 2 4 との詳細は後述する。

[0047] <算出部 1 2 5>

算出部 1 2 5 は、ベクトル定義情報記憶部 2 2 0 からベクトル定義情報を読み出し、読み出したベクトル定義情報とログメッセージとに基づいて、当該ログメッセージに対応する特徴ベクトルを算出する。算出部 1 2 5 は、ログメッセージに、所定の文字列が含まれるか否かに応じて、当該ログメッセージに対応する特徴ベクトルを算出する。

[0048] 図 7 は、ベクトル定義情報記憶部 2 2 0 が記憶するベクトル定義情報の一例を示す図である。ベクトル定義情報は、「特徴ベクトルの基底ベクトルを定義する情報」、および、「基底ベクトルとして定義された文字列の重みを定義する情報」を含む。図 7 において「要素」を示す列および「文字列」を示す列に表される情報が、「特徴ベクトルの基底ベクトルを定義する情報」に相当する。図 7 において「重み」を示す列に表される情報が、「基底ベクトルとして定義された文字列の重みを定義する情報」に相当する。

[0049] まず、「特徴ベクトルの基底ベクトルを定義する情報」について、具体例を用いて説明する。例えば図 7 に示すベクトル定義情報は、要素 1 から要素 1 0 までの 1 0 個の要素を定義する。すなわち、図 7 に示すベクトル定義情報は、特徴ベクトルの次元数（要素数）は 1 0 次元であることを定義する。

- [0050] 図7に示すベクトル定義情報は、特徴ベクトルの1つ目の要素を表す値は、当該特徴ベクトルに対応するログメッセージに「mysqld」という文字列が含まれるか否かに応じて決定されることを示す。
- [0051] 同じく、図7に示すベクトル定義情報は、特徴ベクトルの2つ目の要素を表す値は、当該特徴ベクトルに対応するログメッセージに「innodb」という文字列が含まれるか否かに応じて決定されることを示す。
- [0052] このように、ベクトル定義情報は、特徴ベクトルの特定の要素を表す値が、どの文字列に基づいて決定されるかを定義する情報を含む。
- [0053] 次に、「基底ベクトルとして定義された文字列の重みを定義する情報」について、具体例を用いて説明する。例えば図7に示すベクトル定義情報は、特徴ベクトルにおける「mysqld」に相当する要素を表す値は、当該特徴ベクトルに対応するログメッセージに「mysqld」という文字列が出現する回数と、 $1/67$ という値との積の値により決定されることを示す。
- [0054] 同様に、図7に示すベクトル定義情報は、特徴ベクトルにおける「innodb」に相当する要素を表す値は、当該特徴ベクトルに対応するログメッセージに「innodb」という文字列が出現する回数と、 $1/59$ という値との積の値により決定されることを示す。
- [0055] 算出部125は、ログメッセージ中に、ベクトル定義情報により定義される文字列が含まれるか否かを判定する。ログメッセージ中に、当該文字列が含まれている場合には、算出部125は、ベクトル定義情報が定義する重みの値に応じて、特徴ベクトルにおける、当該文字列に相当する要素を表す値を決定する。なお、算出部125は、ベクトル定義部が定義する文字列が、ログメッセージ中に出現しない場合には、当該文字列に相当する要素を表す値を0とする。
- [0056] 図8は、算出部125が算出した特徴ベクトルと、当該特徴ベクトルに対応するログメッセージとを、関連付けた情報を表す図である。
- [0057] (具体例1) 「130201 09:04:01 mysqld started」というログメッセージから、当該ログメッセージに対応する特徴ベクトルを算出する場合を想定す

る。当該ログメッセージに含まれる複数の文字列の中で、基底ベクトルの要素として定義されている文字列は「mysqld」のみである。したがって、特徴ベクトルは「(1/67, 0, ..., 0)」となる。

[0058] (具体例2) 「130201 09:04:01 InnoDB : started」というログメッセージから、当該ログメッセージに対応する特徴ベクトルを算出する場合を想定する。当該ログメッセージに含まれる複数の文字列の中で、基底ベクトルの要素として定義されている文字列は「innodb」のみである。したがって、特徴ベクトルは「(0, 1/59, 0, ..., 0)」となる。

[0059] (具体例3) 「130201 09:04:01 [note] user/libexec/mysqld」というログメッセージから、当該ログメッセージに対応する特徴ベクトルを算出する場合を想定する。当該ログメッセージに含まれる複数の文字列の中で、基底ベクトルの要素として定義されている文字列は「mysqld」および「libexec」である。したがって、特徴ベクトルは「(1/67, 0, ..., 0, 1/28, 0, ..., 0, ...)」となる。

[0060] 以下、説明の便宜上、文字列抽出部121、文字列集計部122、基底ベクトル定義部123、重み定義部124および算出部125を、特徴ベクトル算出部120とまとめて呼称する場合がある。

[0061] [動作の説明]

第1の実施形態にかかるログ分析システム100の動作を、下記(1)、(2)および(3)の3つの動作に分けて、それぞれ説明する。

- (1) ログ分析システム100がベクトル定義情報を決定する動作、
- (2) ログ分析システム100が分類定義情報を作成する動作、
- (3) ログ分析システム100がログメッセージを分析する動作。

[0062] [(1) ログ分析システム100がベクトル定義情報を決定する動作の説明]

図11は、ログ分析システム100がベクトル定義情報を決定する動作(ステップS101~S107)を説明するフローチャートである。

[0063] ログ収集部110は、分析対象システム900から出力されたログファイ

ル 910 を取得する。ここでログ収集部 110 は、分析対象システム 900 が正常に運用されていた期間におけるログファイル 910 である、正常ログファイル 911 を取得する（ステップ S101）。

[0064] 文字列抽出部 121 は、取得した正常ログファイル 911 を構成するログメッセージの集合から、文字列を抽出する（ステップ S102）。

[0065] 文字列集計部 122 は、文字列抽出部 121 が抽出した文字列毎に、ログメッセージの集合において当該文字列が出現した回数を集計する（ステップ S103）。図 4 は、文字列集計部 122 が集計した情報の一例を示す図である。図 4 に示す情報は、文字列と、当該文字列が出現した回数とを関連付けた情報である。

[0066] 基底ベクトル定義部 123 は、文字列集計部 122 が集計した複数の文字列のうちから、特徴ベクトルの基底ベクトルを定義する文字列を選択する（ステップ S104）。基底ベクトル定義部 123 は、文字列集計部 122 が集計した文字列を特徴ベクトルの基底ベクトルを定義する文字列として選択するか否かを、当該文字列がログメッセージの集合において出現する頻度に応じて決定する。

[0067] ステップ S104 に示した基底ベクトル定義部 123 の動作を、具体例を用いて説明する。基底ベクトル定義部 123 は、変換定義記憶部 210 から、変換閾値定義を読み出す。図 5 は、変換定義記憶部 210 から読み出した変換閾値定義の一例を示す図である。変換閾値定義は、基底ベクトルを定義する文字列を選択する際の判断基準と、その具体的な閾値を定義する。

[0068] 図 5 に示す例では、判断基準が「文字列の出現頻度」、閾値は、「 $0.1N < \sum C_i < 0.6N$ 」である。ここで、 N は集計の対象としたログメッセージの総数を表す値である。 C_i は、あるログメッセージの中に、当該文字列が出現する回数を表す値である。 $\sum C_i$ は、集計の対象としたログメッセージ中に、当該文字列が出現する総回数を表す値である。例えば、上述した「 $0.1N < \sum C_i < 0.6N$ 」という閾値は、200 個のログメッセージを集計の対象とした場合、20 回より多く 120 回より少ない回数出現した文字列を、基底ベクトルを定義

する文字列として選択することを表す。

[0069] したがって、図4に示した文字列のうち、「started」は選択されず、「mysqld」、「innodb」または「libexec」、などが基底ベクトルを定義する文字列として選択される。

[0070] なお、変換閾値定義が定義する閾値の具体的な値は、あらかじめ固定された値でもよいし、ログ収集部110が収集したログファイル910に含まれるログレコードの数に応じて動的に設定されてもよい。あるいは、図示しない運用管理者が定義してもよい。

[0071] 基底ベクトル定義部123は、基底ベクトルを定義する文字列として選択した複数の文字列の順番を一意に特定する（ステップS105）。順番を特定する方法はどのような方法でもよい。基底ベクトル定義部123は例えば、出現頻度が多い順、あるいは辞書順などの方法で文字列の順番を特定する。

[0072] 重み定義部124は、基底ベクトル定義部123が選択したそれぞれの文字列について、重みを定義する（ステップS106）。

[0073] ステップS106に示した重み定義部124の動作を、具体例を用いて説明する。重み定義部124は、変換定義記憶部210から、変換指標定義を読み出す。図6は、変換定義記憶部210から読み出した変換指標定義の一例を示す図である。変換指標定義は、重みを決定する算出方法を定義する。図6に示す変換指標定義は、「文字列の出現頻度の総和の逆数と、係数との積」を重みとして定義することを表す。係数 α を1とした場合、図4に示される情報に基づいて各文字列の重みを定義すると、各文字列の重みは、図7に示す通りとなる。なお、変換指標定義が定義する係数 α は、例えば、基底ベクトルの各要素の重みの総和が1となるように定義されていてもよい。

[0074] 重み定義部124は、基底ベクトル定義部123が選択した、基底ベクトルを定義する文字列および当該文字列の順序、並びに、重み定義部124が定義した重みを、ベクトル定義情報としてベクトル定義情報記憶部220に記憶する（ステップS107）。

[0075] 図7は、基底ベクトル定義部123および重み定義部124が決定したベクトル定義情報の一例を表す図である。

[0076] ここまでで、ログ分析システム100がベクトル定義情報を決定する動作を説明した。

[0077] [(2) ログ分析システム100が分類定義情報を作成する動作の説明]

次に、図12を用いて、ログ分析システム100が分類定義情報を作成する動作を説明する。図12は、ログ分析システム100が分類定義情報を作成する動作(ステップS108~S113)を説明するフローチャートである。

[0078] 算出部125は、ベクトル定義情報記憶部220からベクトル定義情報を読み出す(ステップS108)。

[0079] 算出部125は、読み出したベクトル定義情報と、正常ログファイル911を構成するそれぞれのログメッセージとに基づいて、当該ログメッセージに対応する特徴ベクトルを算出する(ステップS109)。

[0080] 算出部125は、ログメッセージと、当該ログメッセージに対応する特徴ベクトルとを関連付けた情報を出力する(ステップS110)。図8は、ログメッセージと、当該ログメッセージに対応する特徴ベクトルとを関連付けた情報の一例を表す図である。

[0081] 分類指標算出部130は、ログメッセージを、当該ログメッセージに対応する特徴ベクトルに基づいて、複数の集合にクラスタリングする。分類指標算出部130は、クラスタリングの結果、集合を識別する識別子と、当該集合に含まれるログメッセージを識別する識別子とを関連付けた情報を出力する(ステップS111)。

[0082] 図9は、ステップS111に示す動作において分類指標算出部130が出力する情報を説明する図である。図9に示す情報は、例えば、集合ID=1で識別される集合には、ログメッセージID=1、5および8で識別されるログメッセージが属していることを示す。

[0083] ステップS111に示した分類指標算出部130の動作を、具体例を用い

て説明する。分類指標算出部 130 は、算出部 125 が算出した特徴ベクトルに基づいて、特徴ベクトル間の距離を算出する。分類指標算出部 130 は、算出した距離に基づいて、当該特徴ベクトルに対応するログメッセージをクラスタリングし、複数の集合を定義する。

[0084] 特徴ベクトル同士の距離としては例えば、ユークリッド平方距離、ミンコフスキー距離またはマハラノビスの汎距離などを用いることができる。距離を用いてクラスタリングする方法としては、Ward 法、K 平均法、最短距離法、最長距離法または群平均法などが知られている。

[0085] 分類指標算出部 130 は、集合毎に、当該集合に対応する分類指標を算出する（ステップ S112）。分類指標とは、各集合を表現する指標である。分類指標は、各集合に属する特徴ベクトルを用いて算出可能な、値またはベクトルである。発明の理解を容易にするため、以下、分類情報が、ある集合に属する一つまたは複数の特徴ベクトルの平均ベクトルであるとして説明を続ける。平均ベクトルは分類指標の一具体例である。分類指標の他の具体例については後述する。

[0086] 分類指標算出部 130 は、集合 ID と、当該集合に対応する平均ベクトルとを関連付けて、分類定義情報を作成する。分類指標算出部 130 は、作成した分類定義情報を分類定義情報記憶部 230 に記憶する（ステップ S113）。

[0087] 図 10 は、分類指標算出部 130 が作成した分類定義情報の一例を示す図である。図 10 に示す情報は、例えば、集合 ID 1 に対応する平均ベクトルは、 $(1/67, 0, \dots, 0)$ であることを示す。

[0088] ここまでで、ログ分析システム 100 が分類定義情報を作成する動作を説明した。

[0089] なお、ベクトル定義情報および分類定義情報は必ずしも正常ログファイル 911 から作成されとは限らない。例えば、異常なログメッセージの集合からベクトル定義情報および分類定義情報を作成してもよい。

[0090] [(3) ログ分析システム 100 がログメッセージを分析する動作の説明]

図13は、ログ分析システム100がログメッセージを分析する動作を説明するフローチャートである。

- [0091] ログ収集部110は、分析対象システム900から出力されたログファイル910を取得する。ここでログ収集部110は、分析対象のログファイル910である分析対象ログファイル912を取得する（ステップS201）。
- [0092] 特徴ベクトル算出部120は、ベクトル定義情報記憶部220からベクトル定義情報を読み出す（ステップS202）。
- [0093] 特徴ベクトル算出部120は、ベクトル定義情報と、分析対象ログファイル912を構成するログメッセージとに基づいて、ログメッセージに対応する特徴ベクトルを算出する（ステップS203）。
- [0094] なお、ステップS203に示す動作において、分析対象のログメッセージに含まれる文字列が、ベクトル定義情報において定義された要素に含まれない場合がある。このような場合、特徴ベクトル算出部120は、該当する単語を、ベクトル定義情報が定義する要素の末尾に追加することによって（すなわち、図7に示した例においては要素11が追加される）、ベクトル定義情報を拡張するように構成してもよい。その場合、特徴ベクトル算出部120は、ベクトルの次元を該当する単語の数だけ増加させ、その値は0とする。
- [0095] ログ分析部140は、分類定義情報記憶部230から分類定義情報を読み出す（ステップS204）。
- [0096] ログ分析部140は、分類定義情報と、特徴ベクトル算出部120が算出した特徴ベクトルと、に基づいて、当該特徴ベクトルに対応するログメッセージを、分類定義情報が定義する複数の集合のうちいずれかの集合に分類する。あるいはログ分析部140は、ログメッセージを、分類定義情報が定義するいずれの集合にも所属しない異常なログメッセージとして分類する（ステップS205）。
- [0097] ステップS205に示したログ分析部140の動作を、詳細に説明する。ログ分析部140は、分類定義情報が定義する平均ベクトルと、分析対象の

ログメッセージに対応する特徴ベクトルとの距離を算出し、当該距離に基づいて、分析対象のログメッセージを分類する。下記2つの具体例を用いて詳細に説明する。

[0098] (具体例1) 例えば、異常判定対象のログレコード「130201 10:01:23 mysqld started」の特徴ベクトルが、「 $(1/67, 0, \dots, 0)$ 」である場合を想定する。ログ分析部140は、当該特徴ベクトルと、分類定義情報が定義する各々の平均ベクトルとを比較する。ログ分析部140は、分析対象の特徴ベクトルと、平均ベクトルとの距離を算出する。例えば、上述の特徴ベクトルと、図10に示した分類定義情報が定義する平均ベクトルとの距離を算出すると、以下の通りとなる。

集合ID1に対しては0、

集合ID2に対しては $(1/67) * (1/67) + (1/59) * (1/59) + (1/128) * (1/128)$ 、

集合ID3に対しては $(1/28) * (1/28) + (1/264) * (1/264)$ である。

したがって、最も近い集合IDは1である。ここで、記号“*”は乗算を表す。また、記号“/”は除算を表す。以下、これらの記号が表す意味は本願において同様である。

[0099] ログ分析部140は、算出した距離と別途定義された閾値を比較し、算出した距離が閾値以下であれば、分析対象のログレコードが該当する集合に含まれると判定する。上記の例の場合、閾値を $(1/25) * (1/25)$ とすると、距離は0であるため、ログ分析部140は、分析対象のログレコードを、集合ID1の集合に含まれると判定する。

[0100] (具体例2) 例えば、異常判定対象のログレコード「130201 10:01:23 mysqld stopped」の特徴ベクトルが $(1/67, 0, \dots, 1/20)$ である場合を想定する。ログ分析部140は、当該特徴ベクトルと、分類定義情報が定義する各々の平均ベクトルとの距離を算出する。その結果、ログ分析部140は、当該特徴ベクトルともっとも近い平均ベクトルは、集合ID1に対応する平均ベクトルであり、その距離は $(1/20) * (1/20)$ であると算出する。ログ分析部140は更に、算出した距離と閾値とを比較する。距離 $((1/20) * (1/20)) > \text{閾値}$ (

(1/25)*(1/25))であるため、ログ分析部140は、分析対象のログレコードが、いずれの集合にも属さない異常なログレコードであると判定する。

[0101] 出力部150は、ログ分析部140が分析対象のログメッセージを分類した結果を示す情報を出力する（ステップS206）。出力部150は、ログ分析部140から出力された情報を、例えば、ディスプレイに表示する、あるいは、プリンタで帳票として出力するなどの方法で運用管理者に提示する。出力部150は、分類定義情報を出力してもよい。

[0102] [ハードウェア構成の説明]

図14は、第1の実施形態にかかるログ分析システム100のハードウェア構成の一例を示すブロック図である。ログ分析システム100を構成するハードウェアは、CPU（Central Processing Unit）1、メモリ2、記憶装置3、通信インターフェース（I/F）4を備える。ログ分析システム100は、入力装置5または出力装置6を備えていてもよい。ログ分析システム100の機能は、例えばCPU1が、メモリ2に読み出されたコンピュータプログラム（ソフトウェアプログラム、以下単に「プログラム」と記載する）を実行することにより実現される。実行に際して、CPU1は、通信インターフェース4、入力装置5および出力装置6を適宜制御する。

[0103] 尚、本実施形態および後述する各実施形態を例として説明される本発明は、係るプログラムが格納されたコンパクトディスク等の不揮発性の記憶媒体8によっても構成される。記憶媒体8が格納するプログラムは、例えばドライブ装置7により読み出される。

[0104] ログ分析システム100が実行する通信は、例えばOS（Operating System）が提供する機能を使ってアプリケーションプログラムが通信インターフェース4を制御することによって実現される。入力装置5は、例えばキーボード、マウスまたはタッチパネルである。出力装置6は、例えばディスプレイである。ログ分析システム100は、2つ以上の物理的に分離した装置が有線または無線で接続されることによって構成されていてもよい。

[0105] 図14に示すハードウェア構成例は、後述する各実施形態にも適用可能で

ある。なお、ログ分析システム１００は専用の装置であってもよい。なお、ログ分析システム１００およびその各機能ブロックのハードウェア構成は、上述の構成に限定されない。

[0106] 〔第１の実施形態が奏する効果の説明〕

第１の実施形態にかかるログ分析システム１００は、ログメッセージを分析する際に、ログメッセージを適切に分類することができる。その第１の理由は、ログ分析システム１００は、ログメッセージに所定の文字列が含まれているか否かに応じて、当該ログメッセージに対応する特徴ベクトルを算出し、当該特徴ベクトルに基づいて、当該特徴ベクトルに対応するログメッセージを分析するからである。その第２の理由は、ログメッセージを分析する際に用いられる文字列が、分析対象システムから出力されたログメッセージの集合において出現する頻度に応じて選択されるからである。

[0107] 第１の実施形態にかかるログ分析システム１００が奏する効果をより詳細に説明する。分析対象システム９００が正常に運用されている場合、分析対象システム９００は同一のログメッセージを繰り返し出力すると考えられるため、当該ログメッセージを構成する文字列の出現頻度は高くなる。このような、出現頻度の高い文字列は、類似するログメッセージ同士をクラスタリングまたはグルーピングする際に有用な情報である。

[0108] 他方、分析対象システム９００に異常が発生した場合、分析対象システム９００は、正常時とは異なるログメッセージを出力すると考えられる。この場合、当該ログメッセージには、正常なログメッセージに出現しない文字列が含まれていると考えられる。

[0109] ここで、分析対象システム９００が、下記２つのログメッセージを出力した場合を想定する。

[0110] （ログメッセージ１）：2013/10/01 [XXXAPI] auth_method [TAT]00001 [REMOTEHOST]

node00010 [USER_ID] u1011 …、

（ログメッセージ２）：2013/10/01 [YYYAPI] auth_method [TAT]01001 [

REMOTEHOST]

node00034 [USER_ID] u4581 ...。

[0111] 上記ログメッセージ1および2に共通する、[REMOTEHOST]、[TAT]および[USER_ID]などの文字列は出現頻度が高い文字列（言い換えるとログメッセージ内で固定された文字列）である。このような文字列は、ログメッセージ1とログメッセージ2を異なるログメッセージとして分類する上では不要な文字列である。

[0112] 正常なログメッセージと異常なログメッセージとの差異を生む文字列は、例えば、正常なログメッセージに出現しない文字列、特定のログメッセージにのみ出現する文字列、および、ログメッセージ内で変化する文字列である。

[0113] ログ分析システム100は、ログメッセージを分析する際に用いられる文字列を、分析対象システムから出力されたログメッセージの集合において出現する頻度に応じて選択する。かかる構成により、ログ分析システム100は、正常なログメッセージと異常なログメッセージとの差異を生む文字列に着目してログメッセージを分析することができる。これにより、ログ分析システム100は、ログメッセージを適切に分類することができる。

[0114] また、ログ分析システム100は、ログメッセージを、単に正常なログメッセージか異常なログメッセージかという観点で分類するだけでなく、あらかじめ定義された複数の集合のうちいずれかの集合に分類することもできる。

[0115] 近年では、分析対象システム900の安定的な稼働を保証するために、障害が発生した後の対処ではなく、事前に障害などの異常の兆候を検出するという事前対処が求められ始めている。その為には、監視していた異常系のログメッセージだけではなく、正常系のログメッセージを監視する必要がある。ログ分析システム100は、このようなニーズをも満たすことができる。

[0116] <第2の実施形態>

図15は、第2の実施形態にかかるログ分析システム100Aの構成を示

すブロック図である。第2の実施形態にかかるログ分析システム100Aは、第1の実施形態にかかるログ分析システム100が備える出力部150に代えて、処理実行部160を備える。また、第2の実施形態にかかるログ分析システム100Aは、処理定義情報記憶部240を備える。図1に示した構成と実質的に同一の構成については、同様の符号を付与し、説明を省略する。

[0117] 処理定義情報記憶部240は、処理定義情報を記憶する。処理定義情報は、ログ分析部140が分析した結果に基づいて処理実行部160が実行する処理を定義する情報である。

[0118] 図16は、処理定義情報記憶部240が記憶する処理定義情報の一例を説明する図である。図16に示すように、処理定義情報は、条件(condition)とアクション(action)との組によって構成される。条件は、対応するアクションが実行されるための条件を定義する。アクションは、対応する条件が満たされた場合に、処理実行部160が実行する動作を定義する。

[0119] 図16において、ID=1で識別される処理定義情報は、ログ分析部140が、集合ID1に分類されるログメッセージを検出した場合、処理実行部160が、「管理者に通知する」という処理を実行することを定義する。図16において、ID=2で識別される処理定義情報は、ログ分析部140が、異常なログメッセージを検出した場合、処理実行部160が、「管理者に通知し、アラームを出す」という処理を実行することを定義する。図16において、ID=3で識別される処理定義情報は、ログ分析部140が「Connection error」を示すログメッセージを検出した場合、処理実行部160はそれを「通知しない」ことを定義する。図16において、ID=4で識別される処理定義情報は、ログ分析部140が「20:00:01-06:59:59」に出力されたログメッセージを検出した場合、処理実行部160はそれを「07:00:00にまとめて通知」する処理を実行することを定義する。

[0120] なお、処理定義情報は正規表現を用いて記述されていてもよい。また、処理定義情報を構成するアクションは、機械的に実行可能なスクリプトやコマ

ンドでもよい。

[0121] 処理実行部 160 は、ログ分析部 140 が分析した結果に応じて、処理定義情報が定義する処理を実行する。

[0122] 図 17 は、ログ分析システム 100A がログメッセージを分析する動作を説明するフローチャートである。図 17 に示したステップ S201 から S205 までの動作は、図 13 に示したステップ S201 から S205 までの動作と同様であるので、説明を省略する。

[0123] ログ分析部 140 が分析対象のログメッセージを分類した結果に応じて、処理実行部 160 は、処理定義情報が定義する処理を実行する（ステップ S301）。

[0124] [第 2 の実施形態が奏する効果の説明]

第 2 の実施形態にかかるログ分析システム 100A は、ログ分析部 140 が分析した結果に応じて、適切な処理を自動で実行することができる。その理由は、ログ分析システム 100A は、処理実行部 160 を備えるからである。

[0125] <第 3 の実施形態>

図 18 は、第 3 の実施形態にかかるログ分析システム 100B の構成を説明するブロック図である。図 18 に示すように、ログ分析システム 100B は、特徴ベクトル算出手段 120B と、ログ分析手段 140B とを備える。

[0126] 特徴ベクトル算出手段 120B は、分析対象システムから出力されたログメッセージの集合のうち注目するログメッセージに、所定の文字列が含まれているか否かに応じて、当該ログメッセージに対応する特徴ベクトルを算出する。この所定の文字列は、前記ログメッセージの集合に含まれる複数の文字列のうちから、前記ログメッセージの集合において当該文字列が出現する頻度に応じて選択される文字列である。

[0127] ログ分析手段 140B は、特徴ベクトルに基づいて、前記特徴ベクトルに対応する前記ログメッセージを分析する。

[0128] [第 3 の実施形態が奏する効果の説明]

第3の実施形態にかかるログ分析システム100Bは、ログメッセージを適切に分類することができる。

[0129] <その他の実施形態>

ステップS112に示した動作における、分類指標の他の具体例について説明する。

[0130] 分類指標は、集合に属する特徴ベクトルの平均値、最頻値、中央値、あるいは集合の分散、集合に属する特徴ベクトルの個数など、分類指標は、集合に属する一つまたは複数の特徴ベクトルそのものであってもよい。

[0131] 分類指標算出部130は、ある集合に属する一つまたは複数の特徴ベクトルを代表する代表ベクトルを算出し、当該代表ベクトルを分類指標としてもよい。代表ベクトルの各要素を表す値は、例えば、集合に属する複数の特徴ベクトルの各要素を表す値の代表値（すなわち、平均値、最頻値、中央値または分散値など）である。

[0132] ステップS205に示した動作において、ログ分析部140は、必ずしも距離に基づいてログメッセージを分類するとは限らない。例えばログ分析部140は、ログメッセージに対応する特徴ベクトルを集合に追加した場合の、分散や平均ベクトルなど集合の特徴を表現する統計量の変化量やその組み合わせにより算出可能な量などを判定基準として用いてもよい。

[0133] これら判定基準およびその指標は図示しないファイルから読み込むよう構成してもよい。あるいは、ユーザが図示しない入力装置から直接入力するよう構成してもよい。

[0134] 上述した各実施の形態は、適宜組み合わせて実施されることが可能である。

[0135] 各ブロック図に示したブロック分けは、説明の便宜上から表された構成である。各実施形態を例に説明された本発明は、その実装に際して、各ブロック図に示した構成には限定されない。

[0136] また、ログ分析システムの動作の説明においては、複数の動作を順番に説明しているが、その複数の動作の順番は支障のない範囲で変更することがで

きる。また、これら複数の動作は、それぞれ別々のタイミングで実行されるとは限らない。たとえば、ある動作の実行中に他の動作が発生したり、ある動作と他の動作との実行タイミングが部分的に或いは全部において重複したりしてもよい。

[0137] さらに、各動作の説明においては、発明の理解を容易にするため、ある動作が他の動作の契機になるように記載しているが、その記載はある動作と他の動作の関係を限定するものではない。このため、各実施形態を実施するときには、その複数の動作の関係は内容的に支障のない範囲で変更することができる。

[0138] 以上、本発明を実施するための形態について説明したが、上記実施の形態は本発明の理解を容易にするためのものであり、本発明を限定して解釈するためのものではない。本発明はその趣旨を逸脱することなく変更、改良され得ると共に、本発明にはその等価物も含まれる。

産業上の利用可能性

[0139] 本発明は、情報処理システムや物理プラントなどを運用管理する技術に応用することができる。

以上、上述した実施形態を模範的な例として本発明を説明した。しかしながら、本発明は、上述した実施形態には限定されない。即ち、本発明は、本発明のスコープ内において、当業者が理解し得る様々な態様を適用することができる。

この出願は2013年11月12日に出願された日本出願特願2013-233954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。

符号の説明

[0140]	1	CPU
	2	メモリ
	3	記憶装置
	4	通信インターフェース

- 5 入力装置
- 6 出力装置
- 7 ドライブ装置
- 8 記憶媒体
- 1 0 0 ログ分析システム
- 1 0 0 A ログ分析システム
- 1 0 0 B ログ分析システム
- 1 1 0 ログ収集部
- 1 2 0 特徴ベクトル算出部
- 1 2 0 B 特徴ベクトル算出手段
- 1 2 1 文字列抽出部
- 1 2 2 文字列集計部
- 1 2 3 基底ベクトル定義部
- 1 2 4 重み定義部
- 1 2 5 算出部
- 1 3 0 分類指標算出部
- 1 4 0 ログ分析部
- 1 4 0 B ログ分析手段
- 1 5 0 出力部
- 1 6 0 処理実行部
- 2 1 0 変換定義記憶部
- 2 2 0 ベクトル定義情報記憶部
- 2 3 0 分類定義情報記憶部
- 9 0 0 分析対象システム
- 9 1 1 正常ログファイル
- 9 1 2 分析対象ログファイル

請求の範囲

[請求項1] 分析対象システムから出力されたログメッセージの集合のうち注目するログメッセージに対応する特徴ベクトルを、当該ログメッセージが所定の文字列を含むか否かに応じて算出する特徴ベクトル算出手段と、

前記特徴ベクトルに基づいて、前記特徴ベクトルに対応する前記ログメッセージを分析するログ分析手段と、を備え、

前記特徴ベクトル算出手段は、前記所定の文字列を、前記ログメッセージの集合に含まれる複数の文字列のうちから、前記ログメッセージの集合において当該文字列が出現する頻度に応じて選択する、

ログ分析システム。

[請求項2] 前記特徴ベクトルの要素数は2以上であり、前記特徴ベクトルを構成する複数の要素のうち、第1の要素は第1の文字列によって定義されており、第2の要素は第2の文字列によって定義されており、

前記特徴ベクトル算出手段は、前記ログメッセージに対応する特徴ベクトルを算出する際に、前記ログメッセージが、前記特徴ベクトルを構成する複数の要素のうち特定の要素を定義する文字列を含むか否かに応じて、前記特徴ベクトルを構成する前記特定の要素を表す値を決定する、

請求項1に記載のログ分析システム。

[請求項3] 前記特徴ベクトル算出手段は、前記ログメッセージの集合に含まれる複数の文字列のうちから、前記ログメッセージの集合において、第1の閾値以上、かつ、第2の閾値未満の頻度で出現する文字列を、前記所定の文字列として選択する、

請求項1または2に記載のログ分析システム。

[請求項4] 分類指標算出手段を更に備え、

前記分類指標算出手段は、

前記特徴ベクトル算出手段が算出した特徴ベクトルを複数取得し

、

前記複数の特徴ベクトルを、前記特徴ベクトル同士の距離に基づいて、複数の集合にクラスタリングし、

前記集合ごとに、前記集合に属する1つまたは複数の特徴ベクトルに基づいて分類指標を算出し、前記分類指標と前記集合とを関連付け、

前記ログ分析手段は、前記分類指標と、分析対象であるログメッセージに対応する特徴ベクトルとに基づいて、前記特徴ベクトルに対応するログメッセージが、前記分類指標に関連付けされた集合に属するか否かを分析する、

請求項1から3のいずれかに記載のログ分析システム。

[請求項5] 前記ログ分析手段は、前記特徴ベクトルに対応するログメッセージが、前記複数の集合のうちいずれの集合にも分類されない場合、当該ログメッセージを異常なログメッセージとみなす、

請求項4に記載のログ分析システム。

[請求項6] 前記分類指標は、前記集合に属する1つまたは複数の特徴ベクトルに基づいて算出される代表ベクトルであって、前記代表ベクトルを構成する各要素を表す値は、前記1つまたは複数の特徴ベクトルを構成する、対応する要素を表す値の代表値である、

請求項4に記載のログ分析システム。

[請求項7] 前記特徴ベクトル算出手段は、前記ログメッセージに対応する特徴ベクトルを算出する際に、前記ログメッセージが、前記特徴ベクトルを構成する複数の要素のうち特定の要素を定義する文字列を含む場合に、前記特徴ベクトルを構成する前記特定の要素を表す値を、前記ログメッセージに前記の文字列が幾つ含まれるかを表す値と、前記ログメッセージの集合において前記文字列が出現する頻度の逆数に応じた値との、積の値に基づいて決定する、請求項2に記載のログ分析システム。

[請求項8] 処理実行手段を更に備え、

前記処理実行手段は、前記ログ分析手段が分析対象のログメッセージを、前記いずれかの集合に分類するか、あるいは、異常なログメッセージとみなした場合に、前記集合ごとに、あるいは、異常なログメッセージとみなされたことに対応してあらかじめ規定された処理を実行する、

請求項5に記載のログ分析システム。

[請求項9] コンピュータが、

分析対象システムから出力されたログメッセージの集合のうち注目するログメッセージが、所定の文字列を含むか否かに応じて、前記注目するログメッセージに対応する特徴ベクトルを算出し、

前記特徴ベクトルに基づいて、前記特徴ベクトルに対応する前記ログメッセージを分析し、

前記ログメッセージの集合に含まれる複数の文字列のうちから、前記ログメッセージの集合において当該文字列が出現する頻度に応じて前記所定の文字列を選択する、

ログ分析方法。

[請求項10] コンピュータに、

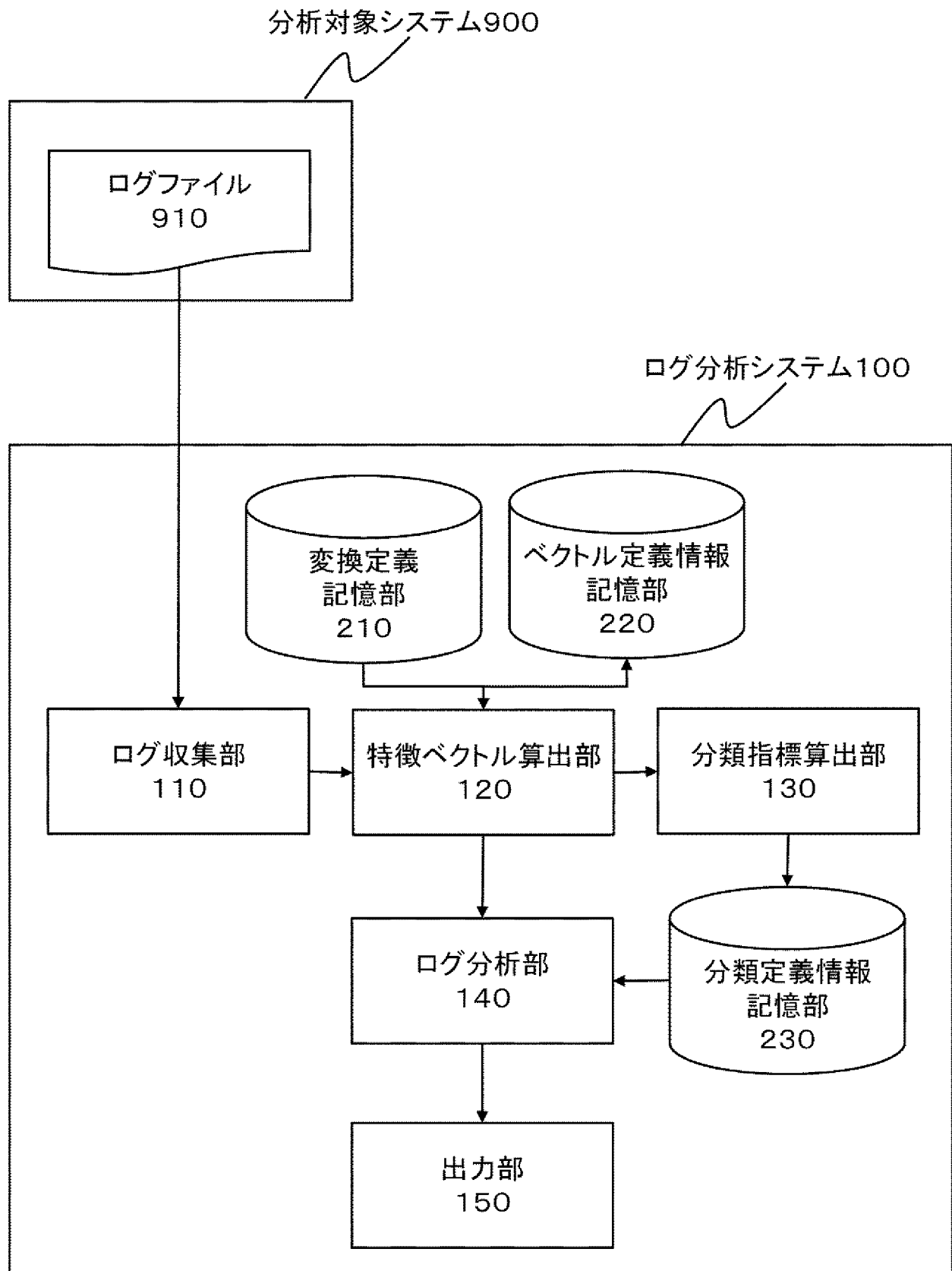
分析対象システムから出力されたログメッセージの集合のうち注目するログメッセージが、所定の文字列を含むか否かに応じて、前記注目するログメッセージに対応する特徴ベクトルを算出する処理と、

前記特徴ベクトルに基づいて、前記特徴ベクトルに対応する前記ログメッセージを分析する処理と、

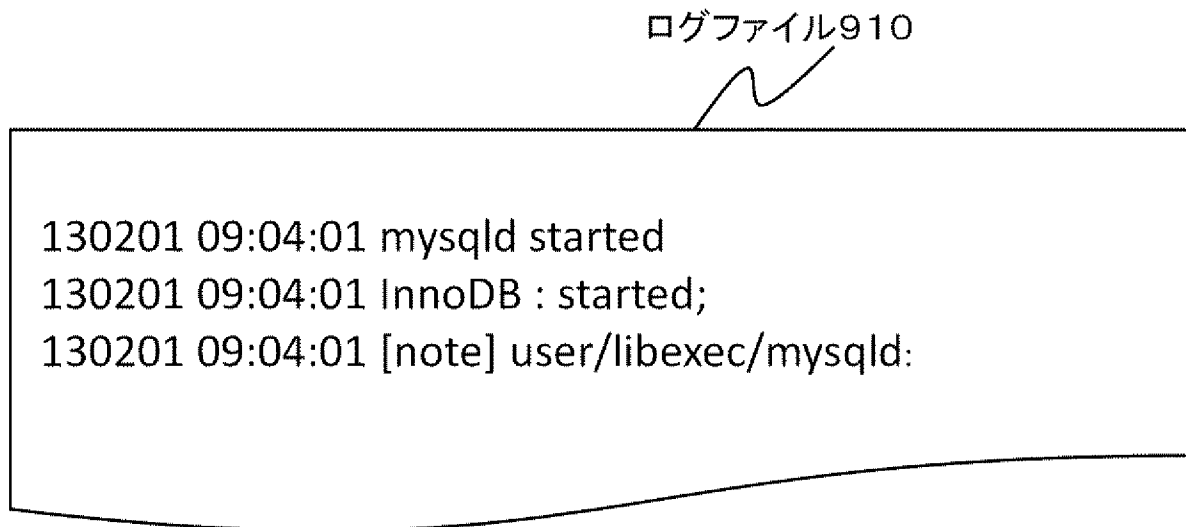
前記ログメッセージの集合に含まれる複数の文字列のうちから、前記ログメッセージの集合において当該文字列が出現する頻度に応じて前記所定の文字列を選択する処理と、

を実行させるプログラムを格納する記憶媒体。

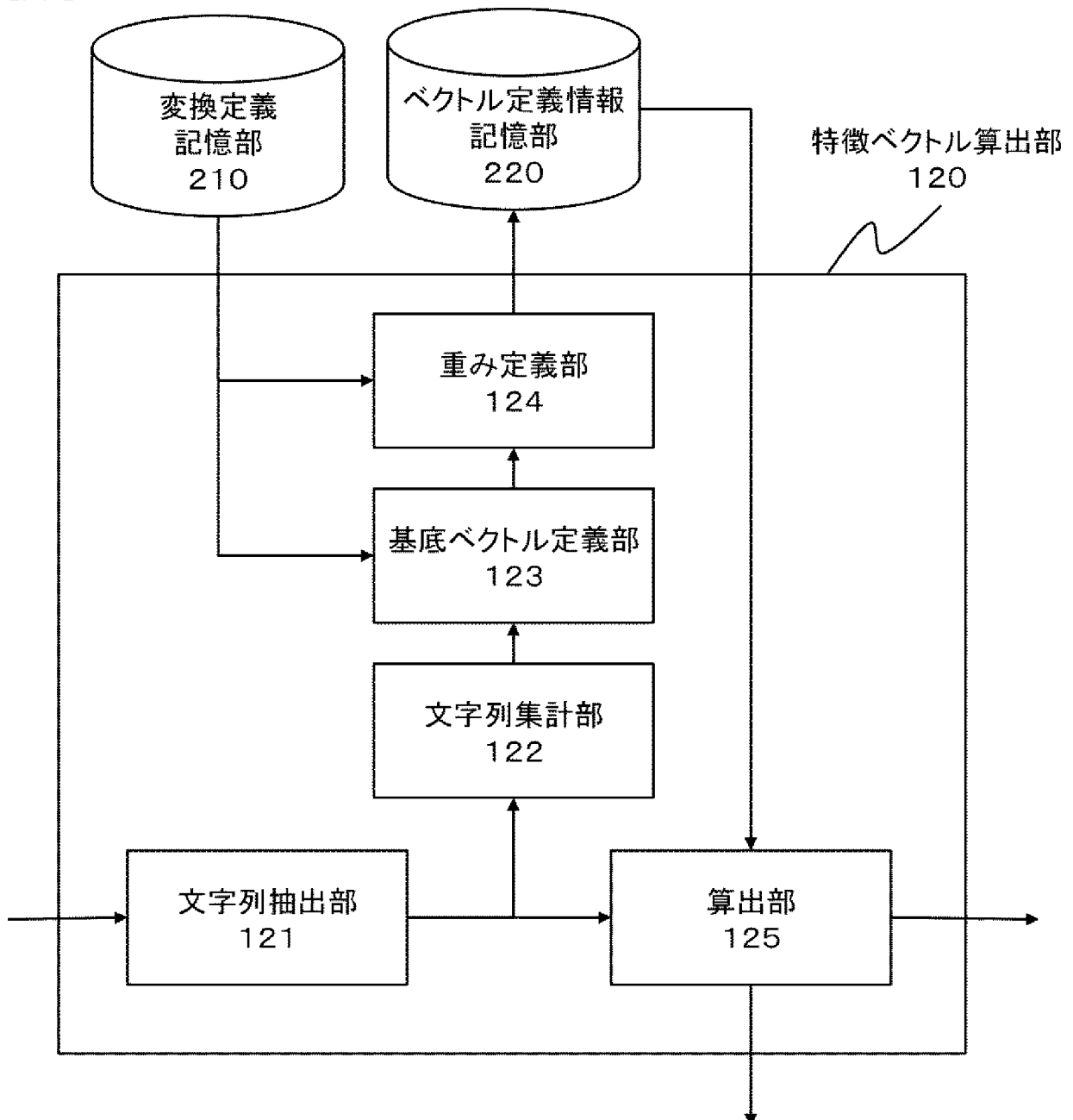
[図1]



[図2]



[図3]



[図4]

文字列	出現回数
started	131
...	...
mysqld	67
innoDB	59
...	...
libexec	28
...	...

[図5]

判断基準	閾値
文字列の 出現頻度	$0.1N < \sum C_i < 0.6N$
...	...

[図6]

指標	算出方法
重み	$\alpha/\sum C_i$
...	...

[図7]

要素	文字列	重み
要素1	mysqld	1/67
要素2	innodb	1/59
...	...	
要素7	libexec	1/28
...	...	
要素10	...	...

[図8]

ログレコード ID	ログレコード	特徴ベクトル
1	130201 09:04:01 mysqld started	(1/67, 0, ..., 0)
2	130201 09:04:01 InnoDB : started;	(0, 1/59, 0, ..., 0)
3	130201 09:04:01 [note] user/libexec/mysqld:	(1/67, 0, ..., 0, 1/28, 0)
...	...	...
...	...	...

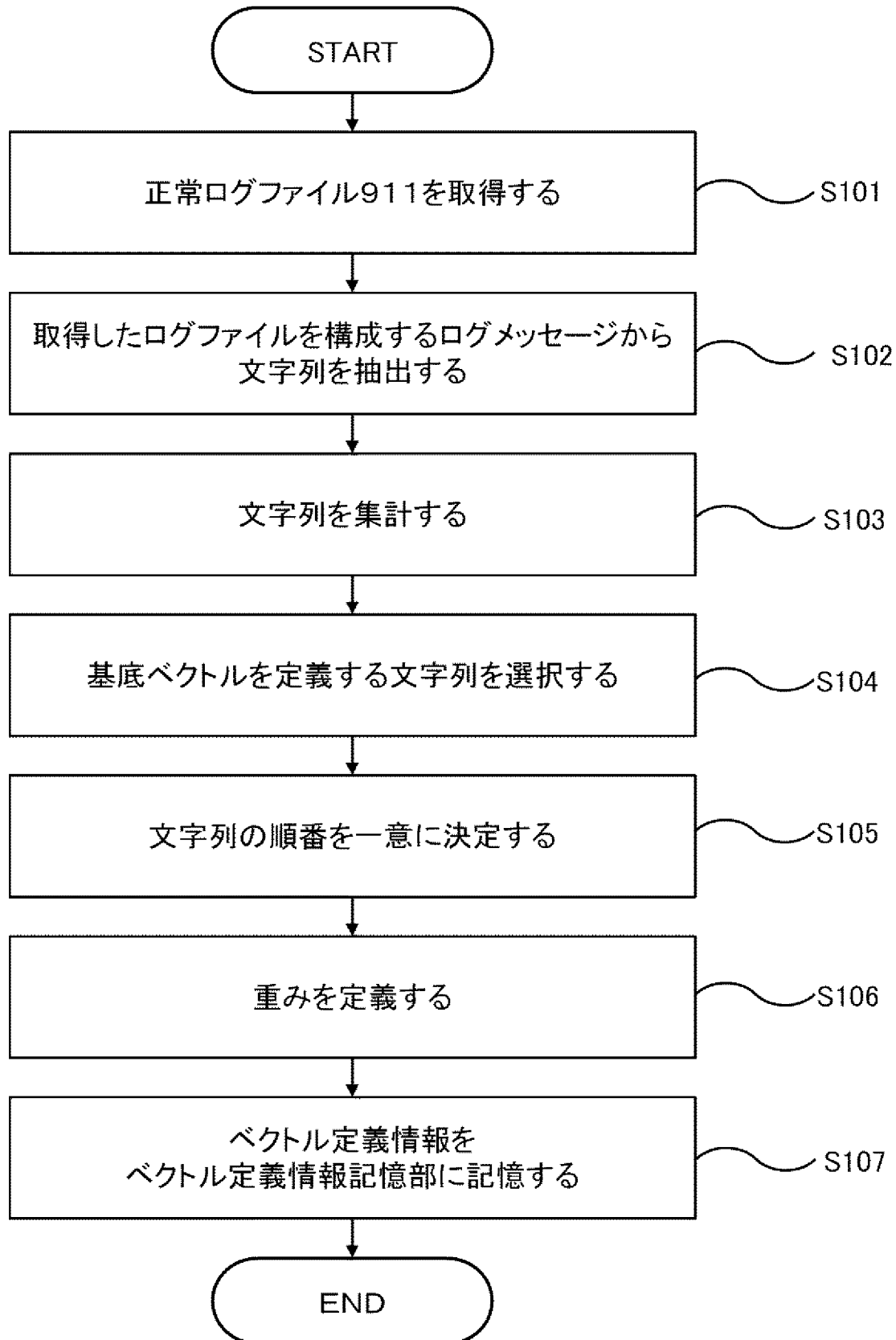
[図9]

集合 ID	ログレコードID
1	1, 5, 8, ...,
2	2, 9, 13, ...,
3	3, 11, 12, ...,
...	...
...	...

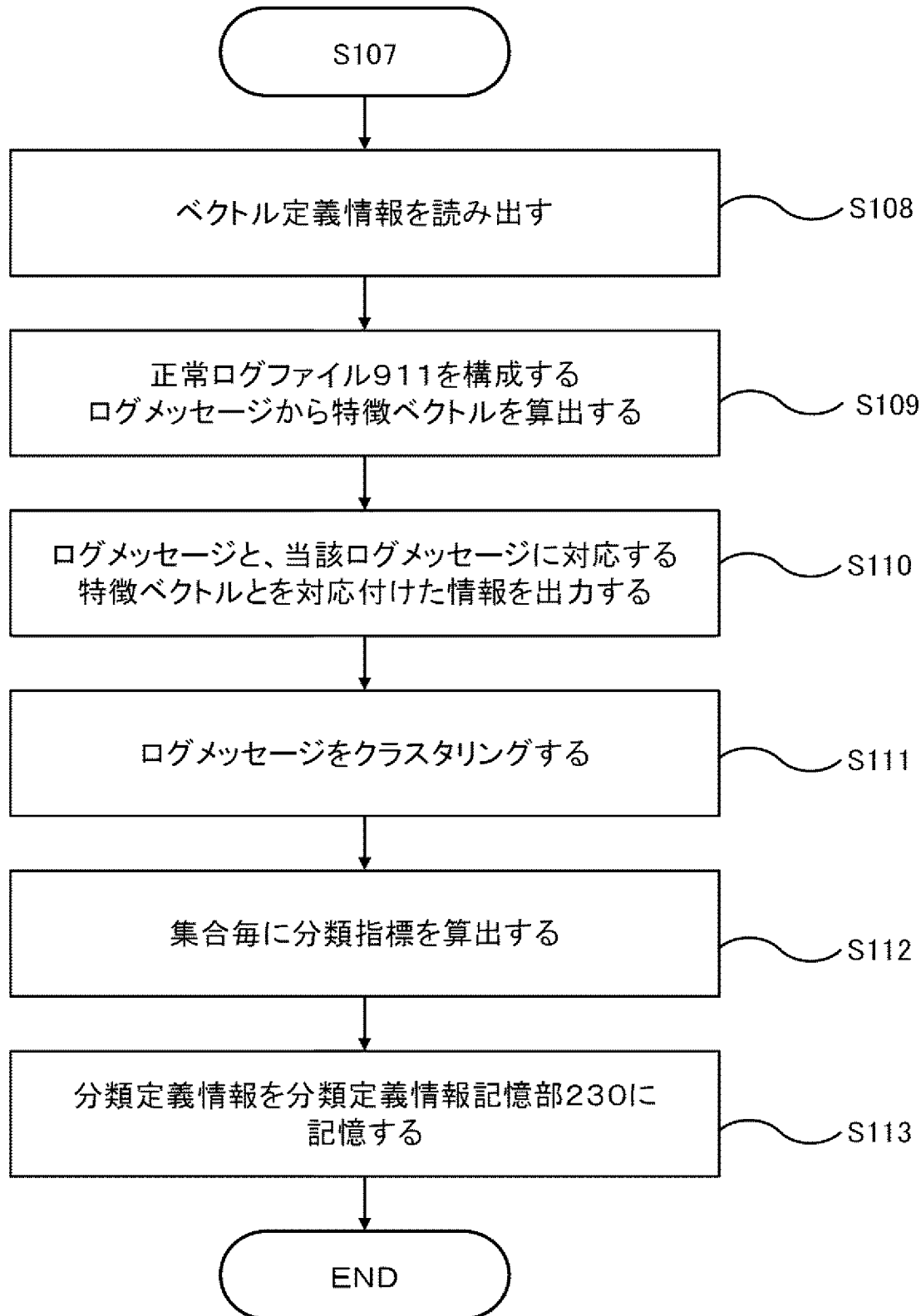
[図10]

集合ID	ログレコードID	分類指標(平均ベクトル)
1	1, 5, 8, ...,	(1/67, 0, ..., 0)
2	2, 9, 13, ...,	(0, 1/59, 0, 1/128, 0, ..., 0)
3	3, 11, 12, ...,	(1/67, 0, ..., 0, 1/28, 0, 1/264)
...	...	...
...	...	...

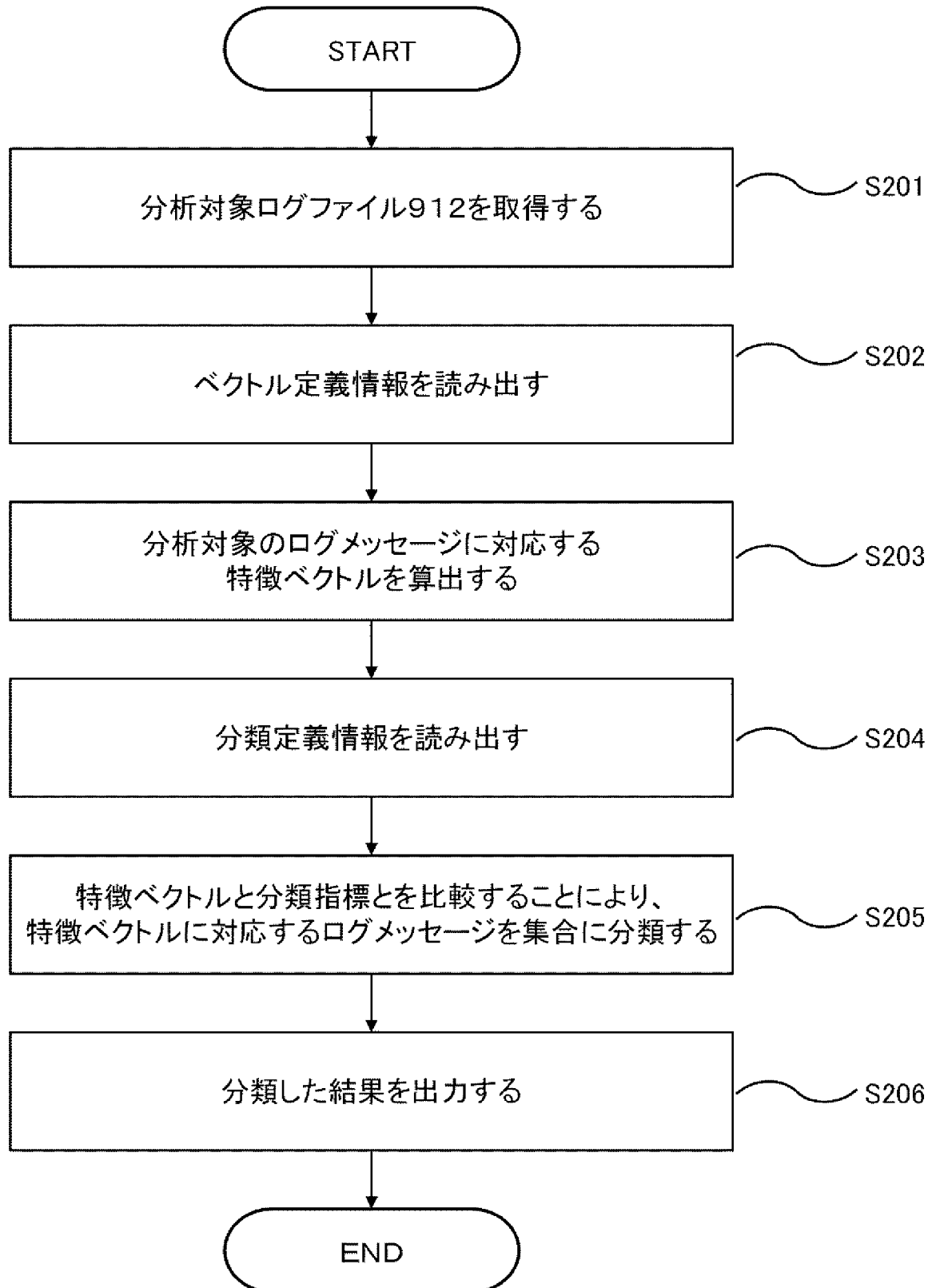
[図11]



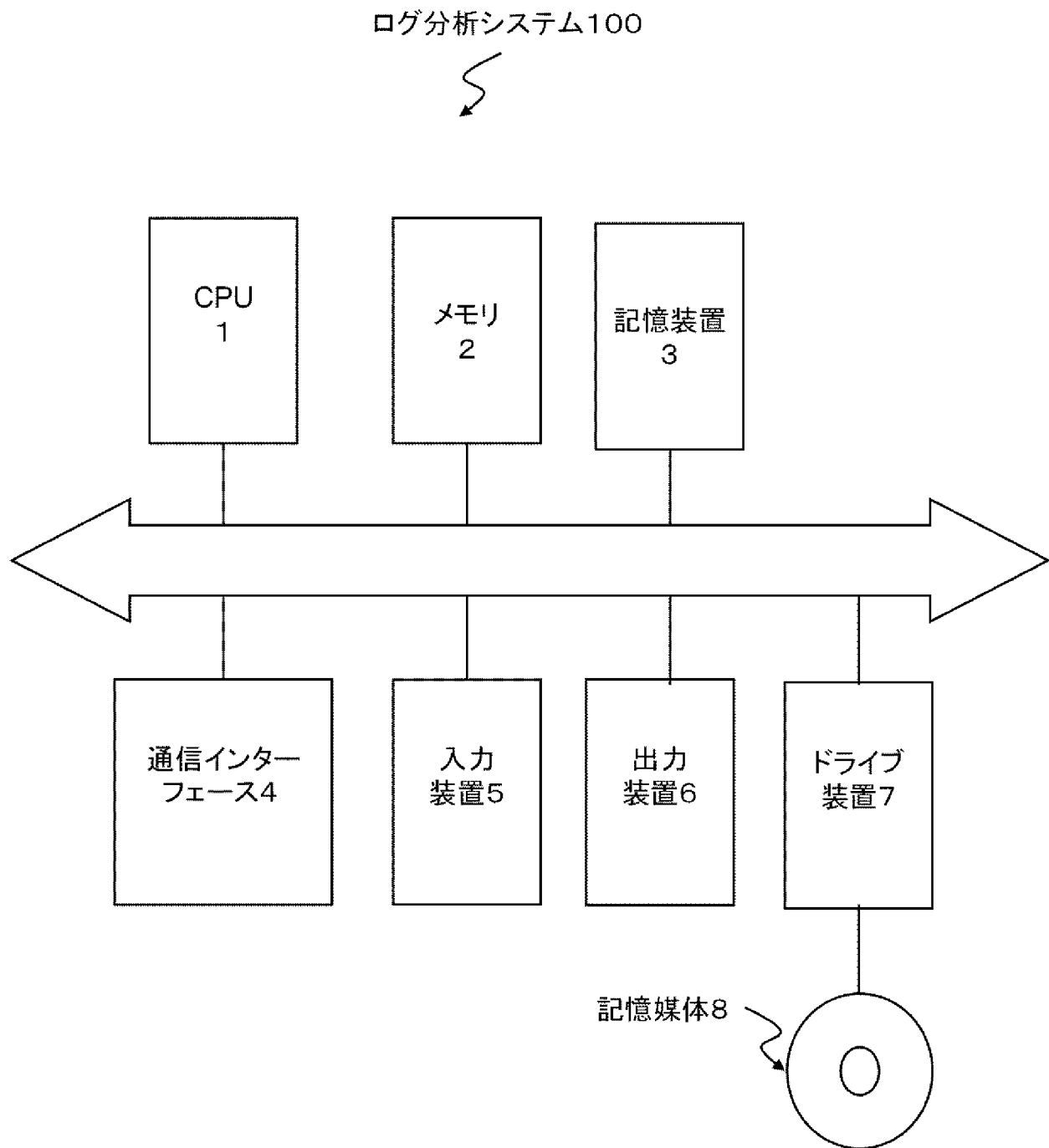
[図12]



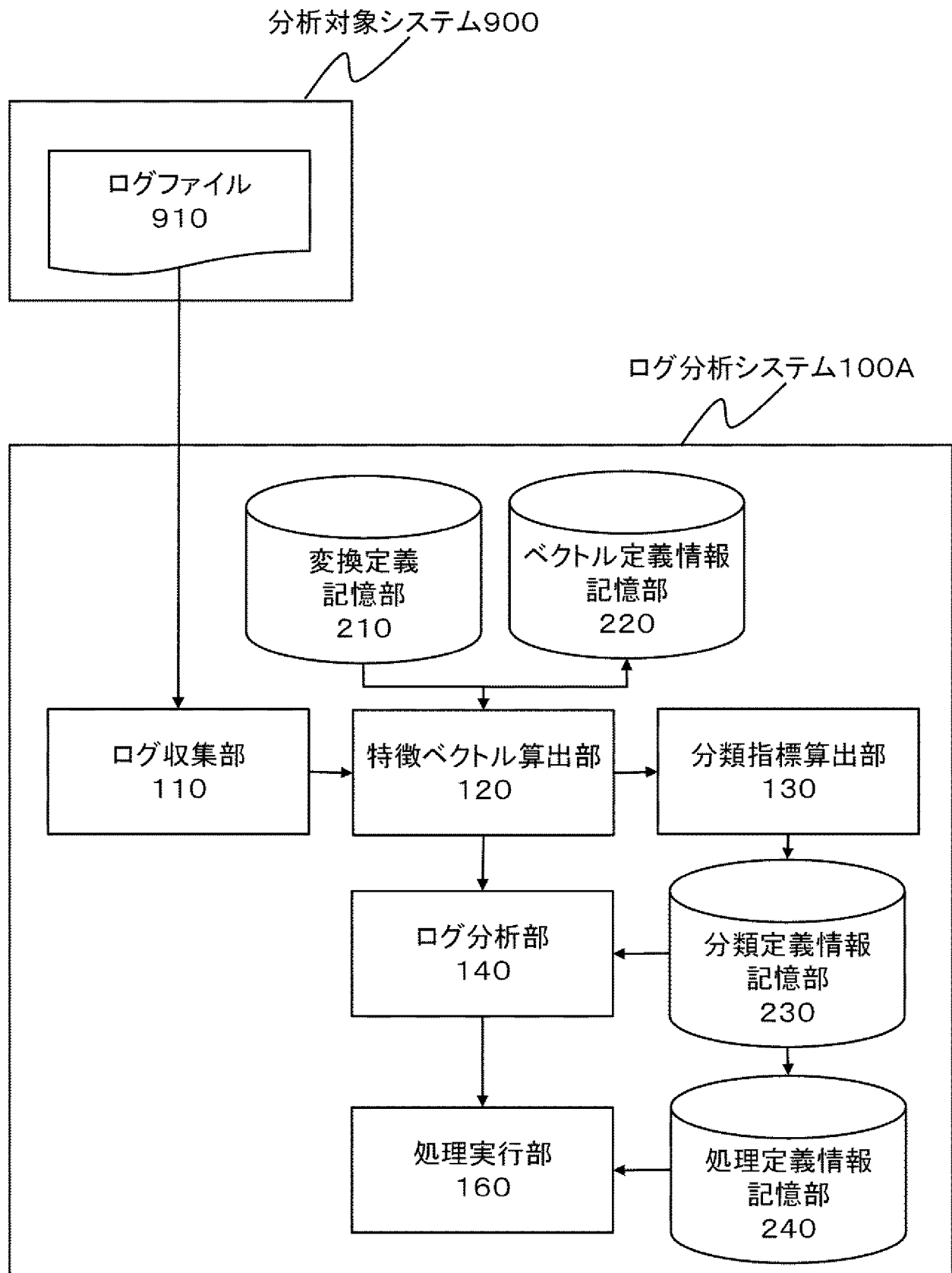
[図13]



[図14]



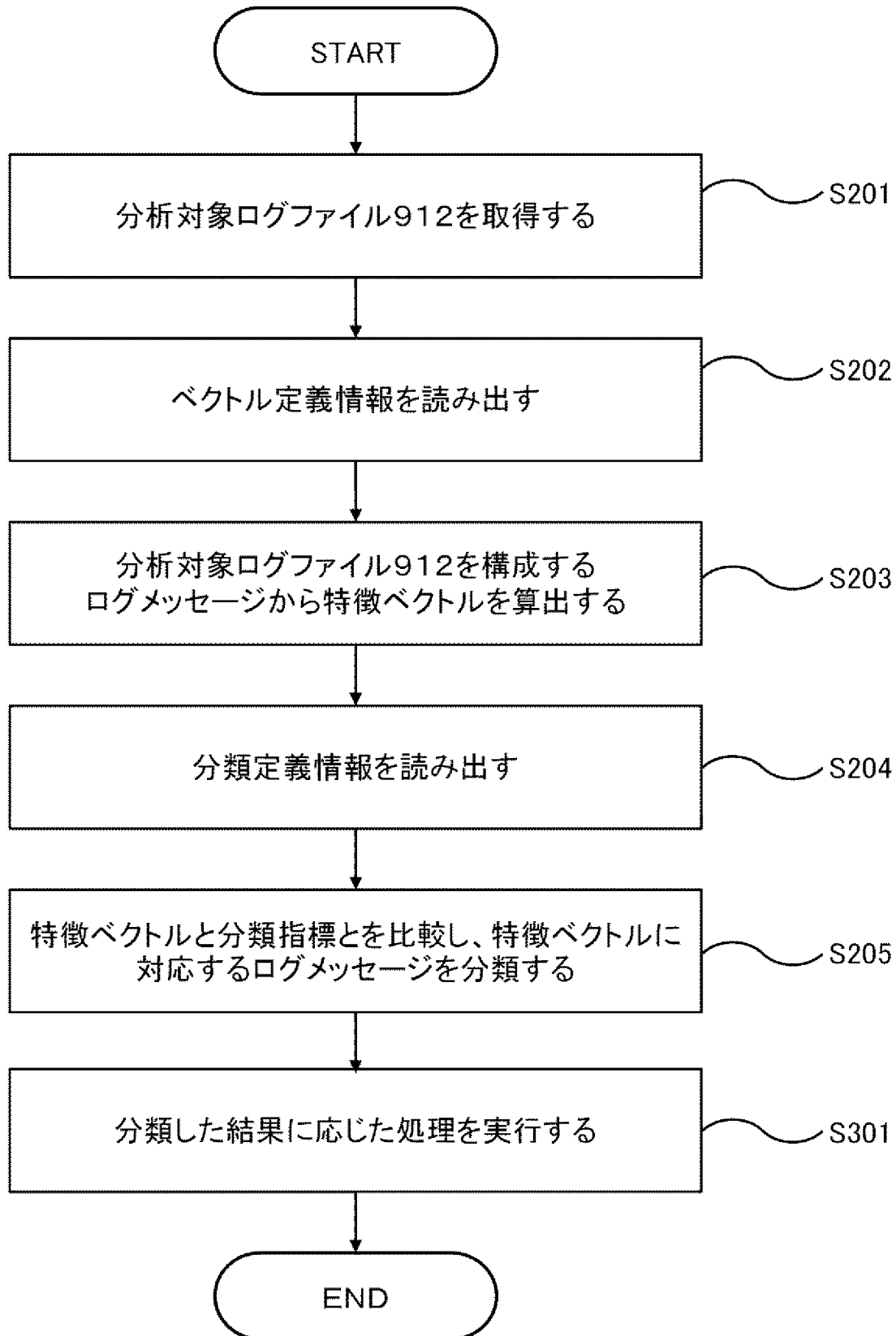
[図15]



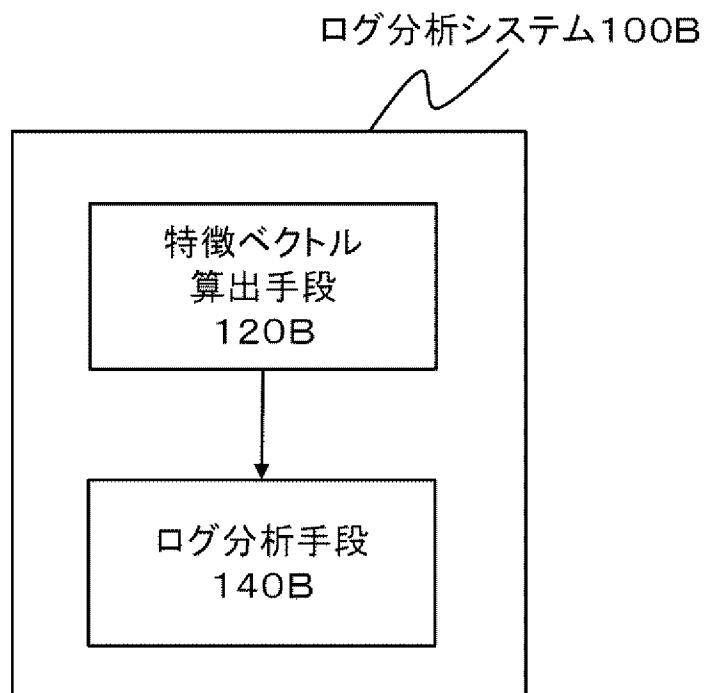
[図16]

ID	条件	アクション
1	集合ID=1	管理者へ通知する
2	異常なログメッセージ	管理者へ通知し、 アラームを出す
3	Connection error*	通知しない
4	20:00:01-06:59:59	07:00:00に管理者へ通知
...	...	...

[図17]



[図18]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2014/005335

A. CLASSIFICATION OF SUBJECT MATTER

G06F11/34(2006.01)i, G06F11/30(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)

G06F11/34, G06F11/30, G06F17/30

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	1996-2014
Kokai Jitsuyo Shinan Koho	1971-2014	Toroku Jitsuyo Shinan Koho	1994-2014

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.
Y	JP 2011-192097 A (Hitachi, Ltd.), 29 September 2011 (29.09.2011), paragraphs [0033] to [0048], [0089]; fig. 3 to 6 (Family: none)	1-10
Y	JP 2004-213626 A (Sony United Kingdom Ltd.), 29 July 2004 (29.07.2004), paragraphs [0029] to [0039] & US 2004/0139105 A1 & GB 2395805 A & EP 1426881 A2	1-10
Y	WO 2010/035412 A1 (NEC Corp.), 01 April 2010 (01.04.2010), paragraph [0073] & US 2011/0137641 A1	7

☐ 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
20 November, 2014 (20.11.14)

Date of mailing of the international search report
02 December, 2014 (02.12.14)

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（IPC））

Int.Cl. G06F11/34(2006.01)i, G06F11/30(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（IPC））

Int.Cl. G06F11/34, G06F11/30, G06F17/30

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	1922-1996年
日本国公開実用新案公報	1971-2014年
日本国実用新案登録公報	1996-2014年
日本国登録実用新案公報	1994-2014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
Y	JP 2011-192097 A（株式会社日立製作所）2011.09.29, 【0033】 －【0048】, 【0089】, 図3-6（ファミリーなし）	1-10
Y	JP 2004-213626 A（ソニー・ユナイテッド・キングダム・リミテッ ド）2004.07.29, 【0029】－【0039】 & US 2004/0139105 A1 & GB 2395805 A & EP 1426881 A2	1-10
Y	WO 2010/035412 A1（日本電気株式会社）2010.04.01, 【0073】 & US 2011/0137641 A1	7

☐ 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

20.11.2014

国際調査報告の発送日

02.12.2014

国際調査機関の名称及びあて先

日本国特許庁（ISA/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）

石川 亮

5 B

3351

電話番号 03-3581-1101 内線 3545