(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2012-173800
(P2012-173800A)
(43)公開日　平成24年9月10日(2012.9.10)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 17/30	(2006.01)	G06F 17/30	350C	5B075
		G06F 17/30	210D	

審査請求 未請求　請求項の数 8　ＯＬ　（全 20 頁）

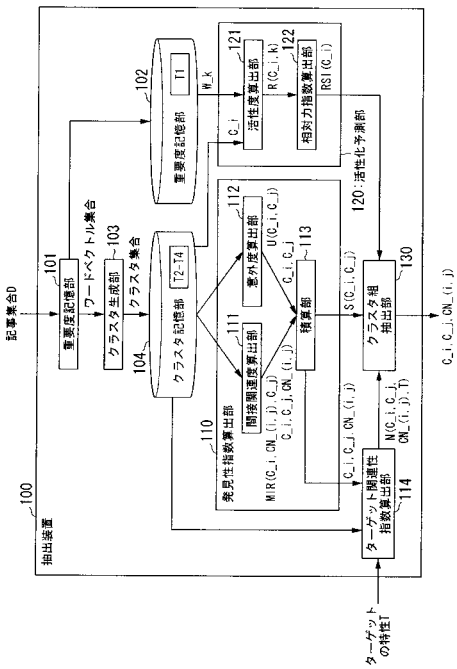
(21)出願番号　特願2011-32415(P2011-32415)
(22)出願日　平成23年2月17日(2011.2.17)

(71)出願人　801000027
学校法人明治大学
東京都千代田区神田駿河台１－１
(74)代理人　100064908
弁理士　志賀　正武
(74)代理人　100106909
弁理士　棚井　澄雄
(74)代理人　100108578
弁理士　高橋　詔男
(74)代理人　100126882
弁理士　五十嵐　光永
(72)発明者　高木　友博
神奈川県川崎市多摩区東三田１－１－１
学校法人明治大学　生田校舎内

Ｆターム(参考)　5B075 NR12　PQ38　PQ75　PR06　QM08

(54)【発明の名称】抽出装置、抽出方法および抽出プログラム

(57)【要約】
【課題】意外性のある概念の組み合わせを提供する。
【解決手段】単語を示す情報と該単語がクラスタに所属している程度である所属度を示す情報とが関連付けられ、単語を示す情報と該単語の位置を示す情報とが関連付けられて記憶されているクラスタ記憶部１０４と、クラスタ記憶部１０４から所属度が所定値以上の単語を示す情報に関連付けられている単語の位置の情報を３つ以上のクラスタ分読み出し、該単語の位置の情報に基づいて、対象となる２つのクラスタ以外の第３のクラスタを経由した該２つのクラスタ間の間接関連度と、該２つのクラスタを組み合わせることの意外度とを反映する発見性指数を算出する発見性指数算出部１１０と、算出された発見性指数に基づいて、前記クラスタの組み合わせを抽出するクラスタ組抽出部１３０を備える。
【選択図】図１



【特許請求の範囲】
【請求項１】
単語を示す情報と該単語がクラスタに所属している程度である所属度を示す情報とが関連付けられ、前記単語を示す情報と該単語の位置を示す情報とが関連付けられて記憶されているクラスタ記憶部と、
前記クラスタ記憶部から所属度が所定値以上の単語を示す情報に関連付けられている単語の位置の情報を３つ以上のクラスタ分読み出し、該単語の位置の情報に基づいて、対象となる２つのクラスタ以外の第３のクラスタを経由した該２つのクラスタ間の間接関連度と、該２つのクラスタを組み合わせることの意外度とを反映する発見性指数を算出する発見性指数算出部と、
前記算出された発見性指数に基づいて、前記クラスタの組み合わせを抽出するクラスタ組抽出部と、
を備えることを特徴とする抽出装置。
【請求項２】
前記クラスタ記憶部から前記クラスタ毎に所属度を示す情報を読み出し、該読み出された所属度を示す情報と、自装置の外部から入力されたターゲットの特性を示す情報とに基づいて、前記対象となる２つのクラスタおよび前記第３のクラスタとターゲットとの関連性を示すターゲット関連性指数を算出するターゲット関連性指数算出部を更に備え、
前記クラスタ組抽出部は、前記発見性指数と前記ターゲット関連性指数とに基づいて、前記クラスタの組み合わせを抽出することを特徴とする請求項１に記載の抽出装置。
【請求項３】
所定の期間毎に、前記単語を示す情報と該単語の重要度を示す情報とが関連付けられて記憶されている重要度記憶部と、
前記重要度記憶部から所定の期間毎に前記単語の重要度を示す情報を読み出し、前記クラスタ記憶部からクラスタ毎に前記所属度を示す情報を読み出し、該単語の重要度を示す情報と該所属度を示す情報とに基づいて、所定の期間毎に各クラスタの活性化を予測する活性化予測部を更に備え、
前記クラスタ組抽出部は、前記活性化予測部による予測により前記クラスタの組み合わせのうち少なくとも１つのクラスタの活性化が予測された場合、前記発見性指数とターゲット関連性指数とに基づいて、前記クラスタの組み合わせを抽出することを特徴とする請求項２に記載の抽出装置。
【請求項４】
前記活性化予測部は、
所定の期間毎に、単語が所定のクラスタへ所属している所属度を示す情報と、前記重要度記憶部から読み出された該期間における前記単語の重要度を示す情報とに基づいて、該クラスタの活性度を算出する活性度算出部と、
前記算出された活性度に基づき、各クラスタの活性度の上昇が期待される度合いである活性度上昇期待値を算出する活性度上昇期待値算出部と、
を備え、
前記算出された活性度と、前記算出された活性度上昇期待値とに基づいて、前記クラスタの活性化を予測することを特徴とする請求項３に記載の抽出装置。
【請求項５】
前記発見性指数算出部は、
前記クラスタ記憶部から前記所属度が所定値以上の単語の位置を示す情報を３つ以上のクラスタ分読み出し、該読み出された単語の位置を示す情報に基づき、対象となる２つのクラスタ以外の第３のクラスタを経由した該２つのクラスタ間の間接関連度を算出する間接関連度算出部と、
前記読み出された単語の位置を示す情報に基づき、前記クラスタの組み合わせの意外度を算出する意外度算出部と、
を備え、

前記間接関連度と前記意外度とを乗じることにより、前記発見性指数を算出することを特徴とする請求項１から請求項４のいずれか１項に記載の抽出装置。
【請求項６】
前記発見性指数は、前記間接関連度と前記意外度が高くなるほど高くなり、
前記クラスタ組抽出部は、前記発見性指数と前記ターゲット指数との重み付き和に基づいて、前記クラスタの組み合わせを抽出することを特徴とする請求項２から請求項５のいずれか１項に記載の抽出装置。
【請求項７】
単語を示す情報と該単語がクラスタに所属している程度である所属度を示す情報とが関連付けられ、前記単語を示す情報と該単語の位置を示す情報とが関連付けられて記憶されているクラスタ記憶部を備える抽出装置が実行する抽出方法であって、
前記クラスタ記憶部から所属度が所定値以上の単語を示す情報に関連付けられている単語の位置の情報を３つ以上のクラスタ分読み出し、該単語の位置の情報に基づいて、対象となる２つのクラスタ以外の第３のクラスタを経由した該２つのクラスタ間の間接関連度と、該２つのクラスタを組み合わせることの意外度とを反映する発見性指数を算出する発見性指数算出手順と、
前記算出された発見性指数に基づいて、前記クラスタの組み合わせを抽出するクラスタ組抽出手順と、
を有することを特徴とする抽出方法。
【請求項８】
単語を示す情報と該単語がクラスタに所属している程度である所属度を示す情報とが関連付けられ、前記単語を示す情報と該単語の位置を示す情報とが関連付けられて記憶されているクラスタ記憶部を備える抽出装置のコンピュータに、
前記クラスタ記憶部から所属度が所定値以上の単語を示す情報に関連付けられている単語の位置の情報を３つ以上のクラスタ分読み出し、該単語の位置の情報に基づいて、対象となる２つのクラスタ以外の第３のクラスタを経由した該２つのクラスタ間の間接関連度と、該２つのクラスタを組み合わせることの意外度とを反映する発見性指数を算出する発見性指数算出ステップと、
前記算出された発見性指数に基づいて、前記クラスタの組み合わせを抽出するクラスタ組抽出ステップと、
を実行させるための抽出プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、抽出装置、抽出方法および抽出プログラムに関する。
【背景技術】
【０００２】
現在、既存の単語を組み合わせることによって作られた造語を新商品の名前に用いることが行われている。その造語が流行するかどうかは、その造語を構成する単語の組み合わせによって変わってくるが、世の中には用いる単語の組み合わせの候補がたくさんあるので、どの単語を組み合わせればよいのかは一見しただけでは分からない。また、あらゆる単語の組み合わせに対して造語が流行するか否かを検証することは難しい。
【０００３】
その問題に対して、非特許文献１では、組み合わせ評価システムがＷＥＢページ上におけるキーワードの登場回数から、単語の組み合わせの斬新さと大衆に受け入られる可能性とを推定し、それによって組み合わせの有効度を定めることが示されている。
【先行技術文献】
【非特許文献】
【０００４】
【非特許文献１】西原陽子、砂山渡、谷内田正彦「有効な組み合わせの発見による創造活

動支援」、電子情報通信学会論文誌　Ｄ－Ｉ，　Ｖｏｌ．Ｊ８７－Ｄ－Ｉ，　Ｎｏ．１０，　ｐｐ．９３９－９４９，　２００４年１０月

【発明の概要】

【発明が解決しようとする課題】

【０００５】

しかしながら、非特許文献１における組み合わせ評価システムは、ＷＥＢページなどの文章に活字として掲載されているキーワードを抽出することはできるが、その文章には活字として掲載されていないが、その文章の一部あるいは全体から捉えられる概念を抽出することができず、意外性のある概念の組み合わせを提供できないという問題があった。

【０００６】

そこで本発明は、上記問題に鑑みてなされたものであり、意外性のある概念の組み合わせを提供することを可能とする抽出装置、抽出方法および抽出プログラムを提供することを課題とする。

【課題を解決するための手段】

【０００７】

上記の課題を解決するために、本発明の一態様である抽出装置は、単語を示す情報と該単語がクラスタに所属している程度である所属度を示す情報とが関連付けられ、前記単語を示す情報と該単語の位置を示す情報とが関連付けられて記憶されているクラスタ記憶部と、前記クラスタ記憶部から所属度が所定値以上の単語を示す情報に関連付けられている単語の位置の情報を３つ以上のクラスタ分読み出し、該単語の位置の情報に基づいて、対象となる２つのクラスタ以外の第３のクラスタを経由した該２つのクラスタ間の間接関連度と、該２つのクラスタを組み合わせることの意外度とを反映する発見性指数を算出する発見性指数算出部と、前記算出された発見性指数に基づいて、前記クラスタの組み合わせを抽出するクラスタ組抽出部と、を備えることを特徴とする。

【０００８】

上記抽出装置は、前記クラスタ記憶部から前記クラスタ毎に所属度を示す情報を読み出し、該読み出された所属度を示す情報と、自装置の外部から入力されたターゲットの特性を示す情報とに基づいて、前記対象となる２つのクラスタおよび前記第３のクラスタとターゲットとの関連性を示すターゲット関連性指数を算出するターゲット関連性指数算出部を更に備え、前記クラスタ組抽出部は、前記発見性指数と前記ターゲット関連性指数とに基づいて、前記クラスタの組み合わせを抽出することを特徴とするものであってもよい。

【０００９】

上記抽出装置は、所定の期間毎に、前記単語を示す情報と該単語の重要度を示す情報とが関連付けられて記憶されている重要度記憶部と、前記重要度記憶部から所定の期間毎に前記単語の重要度を示す情報を読み出し、前記クラスタ記憶部からクラスタ毎に前記所属度を示す情報を読み出し、該単語の重要度を示す情報と該所属度を示す情報とに基づいて、所定の期間毎に各クラスタの活性化を予測する活性化予測部を更に備え、前記クラスタ組抽出部は、前記活性化予測部による予測により前記クラスタの組み合わせのうち少なくとも１つのクラスタの活性化が予測された場合、前記発見性指数とターゲット関連性指数とに基づいて、前記クラスタの組み合わせを抽出することを特徴とするものであってもよい。

【００１０】

上記抽出装置の前記活性化予測部は、所定の期間毎に、単語が所定のクラスタへ所属している所属度を示す情報と、前記重要度記憶部から読み出された該期間における前記単語の重要度を示す情報とに基づいて、該クラスタの活性度を算出する活性度算出部と、前記算出された活性度に基づき、各クラスタの活性度の上昇が期待される度合いである活性度上昇期待値を算出する活性度上昇期待値算出部と、を備え、前記算出された活性度と、前記算出された活性度上昇期待値とに基づいて、前記クラスタの活性化を予測することを特徴とするものであってもよい。

【００１１】

上記抽出装置の前記発見性指数算出部は、前記クラスタ記憶部から前記所属度が所定値以上の単語の位置を示す情報を３つ以上のクラスタ分読み出し、該読み出された単語の位置を示す情報に基づき、対象となる２つのクラスタ以外の第３のクラスタを経由した該２つのクラスタ間の間接関連度を算出する間接関連度算出部と、前記読み出された単語の位置を示す情報に基づき、前記クラスタの組み合わせの意外度を算出する意外度算出部と、を備え、前記間接関連度と前記意外度とを乗じることにより、前記発見性指数を算出することを特徴とするものであってもよい。

【００１２】

上記抽出装置の前記発見性指数は、前記間接関連度と前記意外度が高くなるほど高くなり、前記クラスタ組抽出部は、前記発見性指数と前記ターゲット指数との重み付き和に基づいて、前記クラスタの組み合わせを抽出することを特徴とするものであってもよい。

【００１３】

本発明の一態様である抽出方法は、単語を示す情報と該単語がクラスタに所属している程度である所属度を示す情報とが関連付けられ、前記単語を示す情報と該単語の位置を示す情報とが関連付けられて記憶されているクラスタ記憶部を備える抽出装置が実行する抽出方法であって、前記クラスタ記憶部から所属度が所定値以上の単語を示す情報に関連付けられている単語の位置の情報を３つ以上のクラスタ分読み出し、該単語の位置の情報に基づいて、対象となる２つのクラスタ以外の第３のクラスタを経由した該２つのクラスタ間の間接関連度と、該２つのクラスタを組み合わせることの意外度とを反映する発見性指数を算出する発見性指数算出手順と、前記算出された発見性指数に基づいて、前記クラスタの組み合わせを抽出するクラスタ組抽出手順と、を有することを特徴とする。

【００１４】

本発明の一態様である抽出プログラムは、単語を示す情報と該単語がクラスタに所属している程度である所属度を示す情報とが関連付けられ、前記単語を示す情報と該単語の位置を示す情報とが関連付けられて記憶されているクラスタ記憶部を備える抽出装置のコンピュータに、前記クラスタ記憶部から所属度が所定値以上の単語を示す情報に関連付けられている単語の位置の情報を３つ以上のクラスタ分読み出し、該単語の位置の情報に基づいて、対象となる２つのクラスタ以外の第３のクラスタを経由した該２つのクラスタ間の間接関連度と、該２つのクラスタを組み合わせることの意外度とを反映する発見性指数を算出する発見性指数算出ステップと、前記算出された発見性指数に基づいて、前記クラスタの組み合わせを抽出するクラスタ組抽出ステップと、を実行させるための抽出プログラムである。

【発明の効果】

【００１５】

本発明によれば、意外性のある概念の組み合わせを提供することができる。

【図面の簡単な説明】

【００１６】

【図１】本発明の実施形態における抽出装置のブロック構成図である。

【図２】重要度記憶部に記憶されているワードベクトルテーブルＴ１の一例である。

【図３】クラスタ生成部による処理を説明するための図である。

【図４】活性度の算出方法を説明するための図である。

【図５】本実施形態の抽出装置がクラスタを生成する処理の流れを示したフローチャートである。

【図６】本実施形態の抽出装置がクラスタの組み合わせを抽出する処理の流れを示したフローチャートである。

【発明を実施するための形態】

【００１７】

以下、本発明の実施形態について、図面を参照して詳細に説明する。まず、本発明の実施形態における抽出装置１００の概要について説明する。抽出装置１００は、流行語の重要要素である流行に乗っていることと、新しい驚きがあることとを両立する概念を、その

概念を提供する対象であるターゲット（人）の特性と関連性がある、複数の概念を組み合わせる事によって生成する。これにより、抽出装置１００は、ターゲットの特性に応じて、世間で流行している概念であって、ターゲットにとって意外性がある概念（ヒットコンセプト）を提示することができる。

【００１８】

ここで、概念は、データに出現する語の集合として表される。その特殊な場合として１語による概念も存在する。

組合せ要素となる概念として、２つの概念Ｃ１、Ｃ２をつなぐ役目をする概念Ｃｎが存在する。抽出装置１００は、Ｃ１、Ｃ２、Ｃｎそれぞれの概念を、新聞やウェブ上の時系列データなどから、流行要因（ヒット要因）として定められた特徴を測る測度を測定することで抽出する。

【００１９】

抽出装置１００は、概念Ｃ１と概念Ｃ２の直接の関連度は低いが、概念Ｃｎを経由したＣ１－Ｃｎ－Ｃ２の間接関連度は高くなる組合せを抽出する。例えば、抽出装置１００は、ターゲットがゴルフクラブ（Ｃ１）と関連がある所定の雑誌の読者だとすると、概念Ｃ１、概念Ｃｎ、概念Ｃ２の組み合わせとして、ゴルフクラブ（Ｃ１）、口紅（Ｃ２）、Ｃｎ（プレゼント）を抽出する。一見、ゴルフクラブと口紅の関連度は低いが、プレゼントという概念Ｃｎを経由すると両者の間接関連度は高くなるので、ゴルフクラブ（Ｃ１）と口紅（Ｃ２）の組み合わせを抽出する価値は高い。

【００２０】

さらに、抽出装置１００は、それら概念が対象とする期間において活性化傾向にあることと、それらの概念の少なくとも１つがターゲットの特性と関連があることも概念の抽出の条件とする。例えば、クリスマス時期において、プレゼントという概念の活性化傾向は強くなり、ターゲットとしての所定の雑誌の読者にとってゴルフクラブの関連度は高い。

【００２１】

抽出装置１００は、上記概念Ｃ１、Ｃ２、Ｃｎの組合せを、ターゲットにとって目新しい概念を示す情報とし出力する。これにより、抽出装置１００は、ターゲットに対して、ターゲットと関連している概念（例えば、概念Ｃ１）と、接続概念Ｃｎを介して関連している概念Ｃ２を示す情報を提供することができる。これにより、例えば、ターゲットであるゴルフクラブ（概念Ｃ１）と関連している所定の雑誌の読者に対して、クリスマス時期の流行概念（ヒットコンセプト）として、プレゼント（概念Ｃｎ）のための口紅（概念Ｃ２）特集を提供することができる。

【００２２】

図１は、本発明の実施形態における抽出装置１００のブロック構成図である。抽出装置１００は、重要度算出部１０１と、重要度記憶部１０２と、クラスタ生成部１０３と、クラスタ記憶部１０４と、発見性指数算出部１１０と、ターゲット関連性指数算出部１１４と、活性化予測部１２０と、クラスタ組抽出部１３０とを備える。

また、発見性指数算出部１１０は、間接関連度算出部１１１と、意外度算出部１１２と、積算部１１３とを備える。活性化予測部１２０は、活性度算出部１２１と、相対力指数算出部（活性度上昇期待値算出部）１２２とを備える。

【００２３】

重要度算出部１０１は、自装置の外部から入力された記事集合Ｄを受け付ける。ここで、入力される記事集合Ｄは新聞のような世相を表すドキュメントや雑誌のような市場の特性を表すドキュメントの時系列データである。そして、重要度算出部１０１は、記事集合Ｄから所定期間のドキュメントを一区切りとし、それを時系列順にならべたものを生成する。ここで、一区切りのドキュメントを１つのドキュメント、全期間のドキュメントを全ドキュメントと称する。

【００２４】

重要度算出部１０１は、各期間における単語の重要度を示す情報を算出する。具体的には、例えば、重要度算出部１０１は、期間毎にドキュメント中に注目語が出現した頻度ｔ

ｆを、当該ドキュメント中の総単語数で割ることにより、各期間における各語のｔｆ－ｉｄｆ値を算出する。ここで、ｔｆ－ｉｄｆ値とは、情報検索で一般的に語の重要度として使用されている指標である。

【００２５】

重要度算出部１０１は、このｔｆ－ｉｄｆ値を事前に定められた語順に並べたものであるワードベクトルを当該所定期間毎に算出する。このワードベクトルは、各語のｔｆ－ｉｄｆ値のリストであり、その期間の特徴を表している。重要度算出部１０１は、算出したワードベクトルを示す情報を該単語と関連付けて、期間毎に重要度記憶部１０２のワードベクトルテーブルＴ１に記憶させる。

【００２６】

図２は、重要度記憶部１０２に記憶されているワードベクトルテーブルＴ１の一例である。同図において、上記所定期間を１日と定め、１日毎の単語のｔｆ－ｉｄｆ値が予め決められた単語の順番で示されている。また、各列はワードベクトル（Ｗ＿１、Ｗ＿２、Ｗ＿３、…、Ｗ＿３０）を表している。

このように、このワードベクトルを時系列順に並べることによって、時間順に所定期間毎の記事の特徴が示される。

【００２７】

図１に戻って、重要度算出部１０１は、ワードベクトルの情報の集合（以下、ワードベクトル集合と称する）をクラスタ生成部１０３に出力する。

クラスタ生成部１０３は、重要度算出部１０１から入力されたワードベクトル集合を用いて、単語を所定のまとまりであるクラスタに分類し、クラスタ毎にラベルを付与する。

【００２８】

本実施形態では、概念は何らかの共通性や関連性によって類似の語の集合で表されると仮定する。ここで言う集合とは、その集合の要素であるかどうかの所属度が０または１で決まる通常の集合の場合も、要素の所属度を０から１までの間の任意の値で表すファジィ集合の場合の、両方の可能性がある。

【００２９】

そこで、クラスタ生成部１０３は、所定のクラスタリング方法に従って、記事集合Ｄに出現する単語をクラスタリングする。通常１つのクラスタには数万の単語が含まれ、それぞれの単語はクラスタに所属する値である所属度Ｍｅｍ　Ｃ（ｗ）を有する。ここで、所属度Ｍｅｍ　Ｃ（ｗ）は、単語ｗがクラスタＣに所属する値を表している。この値は、クラスタが対応している概念に所属する程度を意味する。

【００３０】

クラスタリングにはすでに様々な手法が提案されているが、クラスタ生成部１０３は、一例として、ｋ－ｍｅａｎｓ法によって、記事集合Ｄに出現する単語をクラスタリングする。具体的には、クラスタ生成部１０３は、下記式（１）で表される評価値を最小化するクラスタを算出する。ここで、ｋは事前に与えられるものとする。

【００３１】

【数１】

$$\sum_{k=1}^{M}\sum_{i=1}^{N} g_{ik}\,|x_i - v_k|^2 \qquad (1)$$

【００３２】

但し、以下の条件式（２）を満たすものとする。

【００３３】

【数２】

$$\sum_{k=1}^{K} g_{ik} = 1, \quad g_{ik} \in \{0,1\} \qquad (2)$$

【００３４】

ここで、x_iはｉ番目の単語データ（ｉは１からＩまでの整数）で、x_i＝（x_{i1}，x_{i2}）、Ｋはクラスタ数、v_kはｋ番目のクラスタの重心（ｋは１からＫまでの整数）で、v_k＝（v_{k1}，v_{k2}）、g_{ik}はｉ番目のデータのｋ番目のクラスタへの所属度である。

【００３５】

なお、クラスタ生成部１０３は、ｋ－ｍｅａｎｓ法を用いたがこれに限らず、ｆｕｚｚｙ　ｃ－ｍｅａｎｓ法を用いてもよい。その場合、具体的には、クラスタ生成部１０３は、下記式（３）で表される評価値を最小化するクラスタを算出する。ここで、ｋは事前に与えられるものとする。

【００３６】

【数３】

$$\sum_{k=1}^{K} \sum_{i=1}^{I} (g_{ik})^m \, |x_i - v_k|^2 \qquad (3)$$

【００３７】

但し、以下の条件式（４）を満たすものとする。

【００３８】

【数４】

$$\sum_{k=1}^{K} g_{ik} = 1, \quad g_{ik} \in [0,1] \qquad (4)$$

【００３９】

ここで、x_iはｉ番目の単語データ（ｉは１からＩまでの整数）で、x_i＝（x_{i1}，x_{i2}）、Ｋはクラスタ数、v_kはｋ番目のクラスタの重心（ｋは１からＫまでの整数）で、v_i＝（v_{i1}，v_{i2}）、g_{ik}はｉ番目のデータのｋ番目のクラスタへの所属度である。

このように、クラスタ生成部１０３は、ｋ－ｍｅａｎｓ法、ｆｕｚｚｙ　ｃ－ｍｅａｎｓ法のいずれを用いても、要素毎にクラスタに所属する所属度を算出する。

【００４０】

クラスタ生成部１０３は、得られたクラスタ１つずつに１つの概念を割り当てるためにラベルを付与する。具体的には、クラスタ生成部１０３は、クラスタ重心に最も近い語をそのクラスタの代表として、そのクラスタのラベルとする。なお、クラスタ生成部１０３は、クラスタ中の最大の所属度を持つ語をそのクラスタの代表としてそのクラスタのラベルとしてもよい。

【００４１】

図３は、クラスタ生成部１０３による処理を説明するための図である。図３（ａ）は、クラスタ生成部１０３によって生成されるクラスタを説明するための図である。同図において、向かって左側に記事集合Ｄが示されている。向かって右側には、ｘｙの２次元平面上にクラスタの１例が示されている。

【００４２】

その２次元平面上で、クラスタの各要素である各単語は、×印で示されている。３つのクラスタＣ＿１、Ｃ＿２、Ｃ＿３が示されており、各クラスタは円内の×印で示された単

語を含むものとする。クラスタＣ＿１は農産物のラベルが付与されたクラスタであり、その要素にはｐｒｏｃｅｓｓｏｒとｏｒａｎｇｅを含む。一方、クラスタＣ＿２はコンピュータのラベルが付与されたクラスタであり、要素にはｐｒｏｃｅｓｓｏｒ、ｍｅｍｏｒｙを含む。すなわち、ｐｒｏｃｅｓｓｏｒは、食品加工機（フードプロセッサ）という意味でクラスタＣ＿１に所属し、コンピュータのプロセッサの意味でクラスタＣ＿２に所属している。

【００４３】

クラスタＣ＿３は脳のラベルが付与されたクラスタであり、要素にはｍｅｍｏｒｙを含む。すなわち、ｍｅｍｏｒｙは、コンピュータのメモリという意味でクラスタＣ＿２に所属し、脳の記憶という意味でクラスタＣ＿３に所属している。

【００４４】

図３（ｂ）は、クラスタ記憶部１０４に記憶されている概念テーブルＴ２の１例である。概念テーブルＴ２には、図３（ａ）に示されたクラスタを識別する識別情報Ｃ＿ｉ（ｉは正の整数）と、図３（ａ）に示されたクラスタ毎に付与されたラベルを示す情報とが関連付けられている。

【００４５】

図３（ｃ）は、クラスタ記憶部１０４に記憶されている所属度テーブルＴ３の１例である。所属度テーブルＴ３には、図３（ａ）に示された単語を示す情報と、該単語がクラスタに所属している程度である所属度を示す情報とが該クラスタを識別する識別情報Ｃ＿ｉ毎に関連付けられている。

【００４６】

図３（ｄ）は、クラスタ記憶部１０４に記憶されている座標テーブルＴ４の１例である。座標テーブルＴ４には、図３（ａ）に示された単語を示す情報と、該単語の位置を示す情報である座標を示す情報とが関連付けられている。

【００４７】

図１に戻って、クラスタ生成部１０３は、クラスタ識別情報Ｃ＿ｉ（これ以降、ｉはクラスタのインデックスを表す１からｎまでの正の整数）と、クラスタ毎に付与されたラベルを示す情報とを関連付けてクラスタ記憶部１０４に記憶させる。また、クラスタ生成部１０３は、単語を示す情報と、該単語がクラスタに所属している程度である所属度を示す情報とを該クラスタを識別する識別情報Ｃ＿ｉ毎に関連付けてクラスタ記憶部１０４に記憶させる。また、クラスタ生成部１０３は、クラスタ記憶部１０４に、単語を示す情報と当該単語の位置を示す情報とを関連付けて記憶させる。

【００４８】

またクラスタ記憶部１０４には、図３（ｂ）に示されたように、クラスタ生成部１０３による処理の結果、クラスタを識別する識別情報Ｃ＿ｉと、クラスタ毎に付与されたラベルを示す情報とが関連付けられて記憶されている。

またクラスタ記憶部１０４には、図３（ｃ）に示されたように、クラスタ生成部１０３による処理の結果、単語を示す情報と該単語がクラスタに所属している程度である所属度を示す情報とが該クラスタ毎に関連付けられて記憶されている。

【００４９】

クラスタ記憶部１０４には、クラスタ生成部１０３による処理の結果、図３（ｄ）に示されるように、単語を示す情報と、当該単語の位置を示す情報とが関連付けられて記憶されている。ここで、例えば、クラスタ生成部１０３によるクラスタリングにより２次元平面上に、各単語の位置が割り当てられている場合、当該各単語の位置を示す情報は、２次元平面上における座標を示す情報である。

【００５０】

発見性指数算出部１１０は、クラスタ記憶部１０４から異なるクラスタに関連付けられている所属度を示す情報を所定の数（例えば、３つ）のクラスタ分読み出し、当該読み出された所属度を示す情報に基づいて、対象となる２つのクラスタ以外の第３のクラスタを経由した当該２つのクラスタ間の関連度と、該２つのクラスタを組み合わせことの意外度

とを反映する発見性指数を算出する。
ここで、発見性指数は２つのクラスタ同士の直接の関連性が低くなるほど高くなり、該２つのクラスタが残りの第３のクラスタと関連性が高くなるほど高くなる。
【００５１】
間接関連度算出部１１１は、クラスタ記憶部１０４から所属度が所定値以上の単語を示す情報に関連付けられている単語の位置の情報を３つ以上のクラスタ分読み出し、該単語の位置の情報に基づいて、対象となる２つのクラスタ以外の第３のクラスタを経由した該２つのクラスタ間の間接関連度を算出する。一例として、間接関連度算出部１１１は、対象となる２つのクラスタ以外の第３のクラスタを経由したクラスタ間の関連度のうち最大となる最大間接関連度ＭＩＲを算出する。
【００５２】
具体的には、例えば、間接関連度算出部１１１は、クラスタＣ＿ｉとクラスタＣ＿ｊ（これ以降、ｊはクラスタのインデックスを表す１からｎまでの整数）が、接続クラスタＣＮを経由して関連している程度を示す間接関連度のうち、接続クラスタＣＮをＣ＿１からＣ＿ｎまで変化させながら間接関連度を算出し、算出されたｎ個の間接関連度のうち最大となる最大間接関連度ＭＩＲを、下記式（５）を用いて算出する。ここで、接続クラスタＣＮは、Ｃ＿１からＣ＿Ｎまでのクラスタを取りうる。
【００５３】
$$MIR(C\_i, C\_j) = MAX_{CN}\{A(C\_i, CN) \times A(CN, C\_j)\} \quad (5)$$
【００５４】
ここで、MAX_{CN}は、引数である右辺の間接関連度が最大となる接続クラスタＣＮを抽出し、そのときの引数の値を出力する関数で、Ａは第１の引数と第２の引数の関連度を算出する関数である。
なお、間接関連度算出部１１１は、クラスタＣ＿ｉとクラスタＣ＿ｊが、接続クラスタＣＮを経由して関連している程度を示す最大間接関連度ＭＩＲを、下記式（６）を用いて算出してもよい。
【００５５】
$$MIR(C\_i, C\_j) = MAX_{CN}\{A(C\_i, CN) + A(CN, C\_j)\} \quad (6)$$
【００５６】
間接関連度算出部１１１は、式（５）または式（６）の中の関連度Ａを、コサイン類似度を用いて算出する。
【００５７】
一例として、間接関連度算出部１１１がコサイン類似度を用いて関連度Ａを算出する方法について説明する。
ベクトルｘは原点からクラスタＣ＿ｉの重心へのベクトル、ベクトルｙを原点からクラスタＣ＿ｊの重心へのベクトルである。例えば、間接関連度算出部１１１は、以下の式（７）に従って、関連度Ａを算出する。
【００５８】
$$A(C\_i, C\_j) = x \cdot y / (|x| \times |y|) \quad (7)$$
【００５９】
ここで、ｘ・ｙはベクトルｘ、ｙの内積であり、（ｘ１×ｙ１＋ｘ２×ｙ２＋…＋ｘｍ×ｙｍ）で表される（ｍは正の整数）。また、｜ｘ｜はベクトルｘのノルム＝√（ｘ・ｘ）である。式（７）の右辺は、ベクトルｘ、ｙのなす角θの余弦ｃｏｓθを表し、コサイン類似度と呼ばれ、ベクトルの向きの近さ類似性を表す。
【００６０】
なお、間接関連度算出部１１１は、式（５）または式（６）の中の関連度Ａを、ジャカード係数または相互情報量などの方法を用いて算出してもよい。
ジャカード係数を用いる場合には、間接関連度算出部１１１は、Ｃ＿ｉ、Ｃ＿ｊが通常

のクラスタの場合、２つのクラスタＣ＿ｉ、Ｃ＿ｊのどちらかに出現した単語の出現回数によって関連度Ａを算出する。具体的には、間接関連度算出部１１１は、以下の式（８）に従って関連度Ａを算出する。

【００６１】

【数５】

$$A(C\_i,C\_j)=\frac{|C\_i\cap C\_j|}{|C\_i\cup C\_j|} \qquad (8)$$

【００６２】

ここで、｜Ｃ｜はクラスタＣに含まれる要素（単語）数である。この関連度Ａが大きいほど、二つのクラスタの類似性は高い。

クラスタＣ＿ｉ、クラスタＣ＿ｊがｆｕｚｚｙ　ｃ－ｍｅａｎｓ法で算出されたファジィ集合である場合、間接関連度算出部１１１は、x_pをクラスタＣ＿ｉのワードベクトルｘのｐ番目要素（ｐは１からＰまでの整数）、y_qをクラスタＣ＿ｊのワードベクトルｙのｑ番目の要素とすると（ｑは１からＱまでの整数）、クラスタＣ＿ｉ、クラスタＣ＿ｊの関連度を次式（９）で算出する。

【００６３】

【数６】

$$A(C\_i,C\_j)=x\cdot y/\left(\sum_{p=1}^{P}x_p+\sum_{q=1}^{Q}y_q-x\times y\right) \qquad (9)$$

【００６４】

一方、相互情報量を用いる場合には、間接関連度算出部１１１は、下記の式（１０）に従って、クラスタＣ＿ｉ、クラスタＣ＿ｊの相互情報量ＭＩ（Ｃ＿ｉ，Ｃ＿ｊ）を関連度Ａとして算出する。ここで、相互情報量は、ある２つの単語が共起する割合によって求められる関連性の指標である。

【００６５】

【数７】

$$A(C\_i,C\_j)=MI(C\_i,C\_j)=\sum_{p=1}^{P}\sum_{q=1}^{Q}\left[P(x_p,y_q)\times\log\left\{P(x_p,y_q)/(P(x_p)\times P(y_q))\right\}\right] \qquad (10)$$

【００６６】

ここで、x_pはＣ＿ｉのワードベクトルｘのｐ番目の要素、y_qはＣ＿ｊのワードベクトルｙのｑ番目の要素、Ｐ（x_p，y_q）はx_pとy_qの同時出現確率、Ｐ（x_p）、Ｐ（y_q）は、それぞれx_p、y_qの周辺出現確率である。

【００６７】

間接関連度算出部１１１は、クラスタＣ＿ｉとクラスタＣ＿ｊの全ての組み合わせで、最大間接関連度ＭＩＲ（Ｃ＿ｉ，ＣＮ＿（ｉ，ｊ），Ｃ＿ｊ）を算出する。ここで、ＣＮ＿（ｉ，ｊ）は、クラスタＣ＿ｉとクラスタＣ＿ｊとの間接関連度が最大となるときに選択されたクラスタであり、クラスタＣ＿ｉとクラスタＣ＿ｊの組み合わせ毎にクラスタＣ＿１～Ｃ＿Ｎまでの中から選択されたクラスタである。

間接関連度算出部１１１は、算出した全ての最大間接関連度ＭＩＲ（Ｃ＿ｉ，ＣＮ＿（ｉ，ｊ），Ｃ＿ｊ）を示す情報と、その各最大間接関連度ＭＩＲを算出する際に用いたクラスタＣ＿ｉ、ＣＮ＿（ｉ，ｊ）、Ｃ＿ｊの組み合わせを示す情報とを積算部１１３に出力する。

【００６８】

意外度算出部１１２は、クラスタ記憶部１０４から所属度が所定値以上の単語を示す情報を３つ以上のクラスタ分読み出し、該読み出された単語の位置を示す情報に基づき、クラスタの組み合わせの意外度Ｕを算出する。具体的には、例えば、意外度算出部１１２は、式（７）の関連度の式の逆数を意外度として使用し、以下の式に従って、クラスタＣ＿ｉとクラスタＣ＿ｊ間の意外度Ｕ（Ｃ＿ｉ，Ｃ＿ｊ）を算出する。

【００６９】

Ｕ（Ｃ＿ｉ，Ｃ＿ｊ）＝（｜ｘ｜×｜ｙ｜）／ｘ・ｙ　　（１１）

【００７０】

ここで、ベクトルｘは原点からクラスタＣ＿ｉの重心へのベクトル、ベクトルｙを原点からクラスタＣ＿ｊの重心へのベクトルである。

【００７１】

なお、意外度算出部１１２は、ジャッカード係数の逆数（式（７）の右辺の逆数）を用いて、意外度を算出してもよい。その場合、具体的には、意外度算出部１１２は、下記の式（１２）に従って、クラスタＣ＿ｉとクラスタＣ＿ｊ間の意外度Ｕ（Ｃ＿ｉ，Ｃ＿ｊ）を算出する。

【００７２】

【数８】

$$U(C\_i,C\_j)=\frac{|C\_i \cup C\_j|}{|C\_i \cap C\_j|} \qquad (12)$$

【００７３】

ここで、クラスタＣ＿ｉとクラスタＣ＿ｊの関連性が低いほど、意外度Ｕ（Ｃ＿ｉ，Ｃ＿ｊ）は高くなり、両クラスタの組み合わせが意外であることを反映している。

また、意外度算出部１１２は、相互情報量ＭＩの逆数（（式（１０）の右辺の逆数））を用いて、意外度を算出してもよい。その場合、具体的には、意外度算出部１１２は、下記の式（１３）に従って、クラスタＣ＿ｉとクラスタＣ＿ｊ間の意外度Ｕ（Ｃ＿ｉ，Ｃ＿ｊ）を算出する。

【００７４】

Ｕ（Ｃ＿ｉ，Ｃ＿ｊ）＝１／ＭＩ（Ｃ＿ｉ，Ｃ＿ｊ）　　（１３）

【００７５】

意外度算出部１１２は、クラスタＣ＿ｉとクラスタＣ＿ｊの全ての組み合わせで、意外度Ｕ（Ｃ＿ｉ，Ｃ＿ｊ）を算出する。

意外度算出部１１２は、算出した全ての意外度Ｕ（Ｃ＿ｉ，Ｃ＿ｊ）を示す情報と、その各意外度Ｕ（Ｃ＿ｉ，Ｃ＿ｊ）が算出された際に用いられたクラスタＣ＿ｉの識別情報とクラスタＣ＿ｊの識別情報とを積算部１１３に出力する。

【００７６】

続いて、積算部１１３は、最大間接関連度ＭＩＲと意外度Ｕに基づいて、発見性指数を算出する。具体的には、積算部１１３は、対象となる２つのクラスタ（Ｃ＿ｉ、Ｃ＿ｊ）以外の第３のクラスタＣＮを経由した該２つのクラスタ（Ｃ＿ｉ、Ｃ＿ｊ）間の関連度と、該２つのクラスタ（Ｃ＿ｉ、Ｃ＿ｊ）を組み合わせることの意外度とを反映するクラスタ発見性指標Ｓを下記式（１４）に従って、算出する。

【００７７】

Ｓ（Ｃ＿ｉ，Ｃ＿ｊ）＝ＭＩＲ（Ｃ＿ｉ，Ｃ＿ｊ）×Ｕ（Ｃ＿ｉ，Ｃ＿ｊ）　　（１４）

【００７８】

発見性指標Ｓは、クラスタＣ＿ｉとクラスタＣ＿ｊとの間でクラスタＣＮを経由した関連性が必要なこと、また同時にクラスタＣ＿ｉとクラスタＣ＿ｊとの組み合わせに新たな意外性が必要なことを両立させるための指標である。すなわち、発見性指標Ｓは、２つのクラスタ（Ｃ＿ｉ、Ｃ＿ｊ）同士の直接の関連性が低くなるほど高くなり、該２つのクラ

スタが残りの第３のクラスタ（ＣＮ＿（ｉ，ｊ））と関連性が高くなるほど高くなる。

【００７９】

積算部１１３は、クラスタＣ＿ｉとクラスタＣ＿ｊの全ての組み合わせで、発見性指標Ｓを算出し、算出した発見性指標Ｓを示す情報をクラスタ組抽出部１３０に出力する。また、積算部１１３は、クラスタＣ＿ｉを示す情報とクラスタＣ＿ｊを示す情報と接続クラスタＣＮ＿（ｉ，ｊ）を示す情報とをターゲット関連性指数算出部１１４に出力する。

【００８０】

ターゲット関連性指数算出部１１４は、自装置の外部から入力されたターゲットの特性（例えば、ターゲットとなる世相、市場、個人の特性）Ｔを示す情報を受け付ける。また、ターゲット関連性指数算出部１１４は、積算部１１３から入力されたクラスタＣ＿ｉを示す情報とクラスタＣ＿ｊを示す情報と接続クラスタＣＮ＿（ｉ，ｊ）を示す情報とを受け付ける。

【００８１】

ターゲット関連性指数算出部１１４は、クラスタ記憶部１０４からクラスタ（Ｃ＿ｉ、Ｃ＿ｊ、ＣＮ＿（ｉ，ｊ））毎に所属度を示す情報を読み出し、該読み出された所属度を示す情報と、自装置の外部から入力されたターゲットの特性を示す情報Ｔとに基づいて、前記異なる３つのクラスタ（Ｃ＿ｉ、Ｃ＿ｊ、ＣＮ＿（ｉ，ｊ））とターゲットとの関連性を示すターゲット関連性指数Ｎを算出する。

【００８２】

具体的には、例えば、ターゲット関連性指数算出部１１４は、下記の式（１５）に従って、ターゲット関連性指数Ｎを算出する。

【００８３】

Ｎ（Ｃ＿ｉ，Ｃ＿ｊ，ＣＮ＿（ｉ，ｊ），Ｔ）＝ｍｉｎ（Ａ（Ｃ＿ｉ，Ｔ），Ａ（Ｃ＿ｊ，Ｔ），Ａ（ＣＮ＿（ｉ，ｊ），Ｔ））　　（１５）

【００８４】

ターゲット関連性指数算出部１１４は、算出したターゲット関連性指数Ｎを示す情報をクラスタ組抽出部１３０に出力する。

【００８５】

活性度算出部１２１は、各期間のワードベクトルを示す情報を重要度記憶部１０２から読み出し、該読み出された各期間のワードベクトルを示す情報に基づいて、各期間における各クラスタの活性度を算出する。

具体的には、例えば、活性度算出部１２１は、ｋ番目の期間においてｉ番目のクラスタＣ＿ｉの活性度をＲ（Ｃ＿ｉ，ｋ）とすると、下記の式（１６）に従って、活性度を算出する。

【００８６】

Ｒ（Ｃ＿ｉ，ｋ）＝ｓｉｍ（Ｙ＿ｉ，Ｗ＿ｋ）　　（１６）

【００８７】

ここで、Ｙ＿ｉはクラスタＣ＿ｉに所属する単語の所属度から構成される所属度ベクトルであり、Ｗ＿ｋは、ｋ番目（ｋは正の整数）の期間の文書のワードベクトルである。

上記の式（１５）は、活性度算出部１２１は、ｋ番目の期間の文書のワードベクトルＷ＿ｋと、クラスタＣ＿ｉを表す所属度ベクトルＹ＿ｉとの類似度を、そのままそのクラスタＣ＿ｉの活性度として求めるものである。

また、関数ｓｉｍは類似度を表す関数で、コサイン類似度を用いた下記の式（１７）で表される。

【００８８】

ｓｉｍ（Ｙ＿ｉ，Ｗ＿ｋ）＝Ｙ＿ｉ・Ｗ＿ｋ／（｜Ｙ＿ｉ｜×｜Ｗ＿ｋ｜）　　（１７）

【００８９】

図４は、活性度の算出方法を説明するための図である。同図において、所属度ベクトル４０１の各要素は、そのクラスタに属する単語（Ｗｏｒｄ　１～Ｗｏｒｄ　Ｍ）の所属度

が示されている（Ｍは正の整数）。また、ｋ番目の期間の文書のワードベクトル４０２の各要素は、ｋ番目の期間の文書におけるそのクラスタに属する単語（Ｗｏｒｄ　１～Ｗｏｒｄ　Ｍ）のｔｆ－ｉｄｆ値が示されている。

【００９０】

なお、活性度算出部１２１は、関数ｓｉｍとしてジャカード係数を用いてもよい。また、活性度算出部１２１は、下記の式（１８）に従って、クラスタＣ＿ｉの活性度Ｒ（Ｃ＿ｉ）を算出してもよい。

【００９１】

【数９】

$$R(C\_i)=\sum_{p=1}^{P}\sum_{q=1}^{Q}\left\{memC\_i(y_q)\times MI(x_p,y_q)\times tfidf(x_p)\right\} \quad (18)$$

【００９２】

ここで、ｍｅｍ　Ｃ＿ｉ（y_q）は単語y_qのクラスタＣ＿ｉへの所属度である。ＭＩ（x_p，y_q）は，単語x_pと単語y_qとの相互情報量である。ｔｆｉｄｆ（ｘ）はワードベクトル中の単語x_pのｔｆ‐ｉｄｆ値である。

【００９３】

なお、活性度算出部１２１は、各概念に含まれる語すべてを用いて計算する代わりに、ｔｆ‐ｉｄｆ値の高い一定数の上位単語またはｔｆ‐ｉｄｆ値が所定の値を超えた単語のｔｆ‐ｉｄｆ値から構成されるワードベクトルに基づいて活性度を算出してもよい。これにより、活性度算出部１２１は、計算回数を少なくすることができるので、計算に係る時間を短縮することができる。

【００９４】

活性度算出部１２１は、算出した各期間のクラスタＣ＿ｉの活性度Ｒ（Ｃ＿ｉ，ｋ）を示す情報を相対力指数算出部１２２に出力する。

相対力指数算出部１２２は、活性度算出部１２１から入力された各期間のクラスタＣ＿ｉの活性度Ｒ（Ｃ＿ｉ，ｋ）に基づいて、それぞれのクラスタの活性度の時間的変化に注目し、世の中一般やターゲット市場さらには個人で、各クラスタの活性度の上昇が期待される度合い（活性度上昇期待値）を算出する。

【００９５】

具体的には、例えば、相対力指数算出部１２２は、活性度上昇期待値の一例として、相対力指数ＲＳＩ（Ｃ＿ｉ）を算出する。ここで、相対力指数（ＲＳＩ）とは、過去の値の動きに対する上昇幅の割合を求めたもので、一般にＲＳＩ値が３０を切ると、上昇傾向になると言われている。相対力指数算出部１２２は相対力指数（ＲＳＩ）を算出する際に、例えば１カ月あるいは１日のような所定の長さのサンプリング期間を設けて、そのサンプリング期間内の活性度の上昇値と下降値から、相対力指数（ＲＳＩ）を算出する。

例えば、相対力指数算出部１２２は、下記の式（１９）に従って、相対力指数（ＲＳＩ）を算出する。

【００９６】

ＲＳＩ＝ｕ／（ｕ＋ｄ）×１００　　（１９）

【００９７】

ここで、ｕは所定のサンプリング期間の活性度の上昇値の合計、ｄは所定のサンプリング期間の活性度の下降値の合計である。

なお、相対力指数算出部１２２は、活性度上昇期待値として相対力指数ＲＳＩを用いたが、これに限らず、他の経済指標を用いてもよい。

【００９８】

そして、活性化予測部１２０は、算出された活性度と、算出された活性度上昇期待値とに基づいて、クラスタの活性化を予測する。

具体的には、活性化予測部１２０は、上記の３０という値を一般化して閾値Ｌとし、上

昇を予測する条件を下記の２つとする。１つ目は、（ｉ）過去の一定期間の間に相対力指数（ＲＳＩ）が閾値Ｌを下回ったことがあること、２つ目は、（ｉｉ）現在の活性値Ｒが上限Ｒｕ、下限ＲＬの間にあることである。活性化予測部１２０は、これら２つの条件を満たしたときに、これからのクラスタの活性化を予測し、それ以外の場合、これからクラスタが活性化しないと予測する。

【００９９】

活性化予測部１２０は、予測結果を示す情報をクラスタ組抽出部１３０に出力する。

クラスタ組抽出部１３０は、積算部１１３から発見性指標Ｓを示す情報を、ターゲット関連性指数算出部１１４からターゲット関連性指数Ｎを示す情報を、活性化予測部１２０から予測結果を示す情報を受け取る。

【０１００】

クラスタ組抽出部１３０は、活性化予測部１２０による予測により前記クラスタの組み合わせのうち少なくとも１つのクラスタの活性化が予測された場合、発見性指数Ｓとターゲット関連性指数Ｎとに基づいて、クラスタの組み合わせを抽出する。

具体的には、クラスタ組抽出部１３０は、下記の３つの条件に基づいて、クラスタの組み合わせ（Ｃ＿ｉ、Ｃ＿ｊ、ＣＮ（ｉ，ｊ））を抽出する。

【０１０１】

（１）新規発見性指数Ｓの条件として、クラスタの組Ｃ＿ｉ、Ｃ＿ｊ、ＣＮ（ｉ，ｊ）の発見性指標Ｓが所定の値以上であること、

（２）活性化予測の条件として、クラスタＣ＿ｉ、クラスタＣ＿ｊ、クラスタＣＮ（ｉ，ｊ）のいずれかの相対力指数（ＲＳＩ）と活性度Ｒが、それぞれ上述のクラスタの活性化予測条件（ｉ）および（ｉｉ）を満足していること、

（３）ターゲット関連性指数Ｎの条件として、クラスタの組Ｃ＿ｉ、Ｃ＿ｊ、ＣＮ（ｉ，ｊ）のいずれかが、ターゲットの特性Ｔと所定の値以上の関連度を持つことである。

【０１０２】

例えば、クラスタ組抽出部１３０は、あるターゲットの特性Ｔが存在した時、特性Ｔにとっての最適なクラスタの組み合わせ（Ｃ＿ｉ、Ｃ＿ｊ、ＣＮ（ｉ，ｊ））を、下記の式（２０）から算出する。

【０１０３】

ａｒｇ　ｍａｘ｛ａＳ（Ｃ＿ｉ，Ｃ＿ｊ，ＣＮ（ｉ，ｊ））＋ｂＮ（Ｃ＿ｉ，Ｃ＿ｊ，ＣＮ（ｉ，ｊ），Ｔ）｝　　（２０）

【０１０４】

ここで、ａ、ｂはＳ、Ｎに対する重みを表す係数であり、ａｒｇ　ｍａｘは、引数が最大となる値を求める関数である。この式（１８）により、クラスタ組抽出部１３０は、引数の値が最大となるクラスタの組み合わせを抽出することができる。ただし，Ｃ＿ｉ、Ｃ＿ｊ、ＣＮ（ｉ，ｊ）のうちいずれかの相対力指数（ＲＳＩ）と活性度Ｒが、それぞれクラスタの活性化予測条件（ｉ）および（ｉｉ）を満足していることとする。

【０１０５】

なお、本実施形態では、クラスタ組抽出部１３０は、一例として、式（２０）の引数が最大となるクラスタの組み合わせを１つ抽出したが、これに限ったものではない。クラスタ組抽出部１３０は、式（２０）の引数の値が所定の値以上となる１つ以上のクラスタの組み合わせすべてを抽出してもよい。また、クラスタ組抽出部１３０は、式（２０）の引数の値が高いほうからトップＭ（Ｍは正の整数）のクラスタの組み合わせすべてを抽出してもよい。

【０１０６】

そして、クラスタ組抽出部１３０は、抽出したクラスタの組み合わせを構成するクラスタＣ＿ｉを示す情報とクラスタＣ＿ｊを示す情報とクラスタＣＮ＿（ｉ，ｊ）を示す情報とを自装置の外部に出力する。

なお、クラスタ組抽出部１３０は、抽出したクラスタの組み合わせを構成する各クラスタに関連付けられたラベルをそれぞれクラスタ記憶部１０４のテーブルＴ２から読み出し

、読み出した各ラベルを示す情報をヒットコンセプトの組み合わせを示す情報として自装置の外部に出力してもよい。

【０１０７】

図５は、本実施形態の抽出装置１００がクラスタを生成する処理の流れを示したフローチャートである。まず、重要度算出部１０１は、所定期間毎の一区切りのドキュメント中に掲載された各単語のｔｆ－ｉｄｆ値の算出する（ステップＳ１０１）。次に、重要度算出部１０１は、所定期間毎に、各単語のｔｆ－ｉｄｆ値が予め決められた単語順に並べられたワードベクトルを算出する（ステップＳ１０２）。

【０１０８】

重要度算出部１０１は、全期間のドキュメントでワードベクトルを算出したか判定する（ステップＳ１０３）。重要度算出部１０１は、全期間のドキュメントでワードベクトルを算出していない場合（ステップＳ１０３　ＮＯ）、ステップＳ１０１の処理に戻る。一方、重要度算出部１０１が、全期間のドキュメントでワードベクトルを算出した場合（ステップＳ１０３　ＹＥＳ）、クラスタ生成部１０３は、クラスタを生成する（ステップＳ１０４）。

【０１０９】

次に、クラスタ生成部１０３は、単語毎にクラスタへの所属度を算出する（ステップＳ１０５）。次に、クラスタ生成部１０３は、クラスタ毎にクラスタのラベルを抽出する（ステップＳ１０６）。次に、クラスタ生成部１０３は、クラスタの識別情報とクラスタのラベルを示す情報とを関連付けて、クラスタ記憶部１０４に記憶させる（ステップＳ１０７）。次に、クラスタ生成部１０３は、単語を示す情報と各クラスタへの所属度を示す情報とをクラスタ毎に関連付けてクラスタ記憶部１０４に記憶させる（ステップＳ１０８）。以上で、本フローチャートの処理を終了する。

【０１１０】

以上により、抽出装置１００は、記事集合Ｄから所定期間毎の一区切りのドキュメント中に掲載された各単語の重要度を算出することができる。また、抽出装置１００は、記事集合Ｄからクラスタを生成することができる。

【０１１１】

図６は、本実施形態の抽出装置１００がクラスタの組み合わせを抽出する処理の流れを示したフローチャートである。まず、間接関連度算出部１１１は、最大間接関連度ＭＩＲを算出する（ステップＳ２０１）。次に、間接関連度算出部１１１は、全てのクラスタの組み合わせで最大間接関連度ＭＩＲを算出したか否か判定する（ステップＳ２０２）。間接関連度算出部１１１は、全てのクラスタの組み合わせで最大間接関連度ＭＩＲを算出していない場合（ステップＳ２０２　ＮＯ）、ステップＳ２０１の処理に戻る。

【０１１２】

一方、間接関連度算出部１１１が全てのクラスタの組み合わせで最大間接関連度ＭＩＲを算出した場合（ステップＳ２０２　ＹＥＳ）、意外度算出部１１２は、意外度Ｕを算出する（ステップＳ２０３）。次に、意外度算出部１１２は、全てのクラスタの組み合わせで意外度Ｕを算出したか否か判定する（ステップＳ２０４）。意外度算出部１１２は、全てのクラスタの組み合わせで意外度Ｕを算出していない場合（ステップＳ２０４　ＮＯ）、ステップＳ２０３の処理に戻る。

【０１１３】

一方、意外度算出部１１２が全てのクラスタの組み合わせで意外度Ｕを算出した場合（ステップＳ２０４　ＹＥＳ）、積算部１１３は、発見性指標を算出する（ステップＳ２０５）。次に、積算部１１３は、全期間のドキュメントで発見性指標を算出したか否か判定する（ステップＳ２０６）。積算部１１３は、全期間のドキュメントで発見性指標を算出していない場合（ステップＳ２０６　ＮＯ）、ステップＳ２０１の処理に戻る。

【０１１４】

一方、積算部１１３が全期間のドキュメントで発見性指標を算出した場合（ステップＳ２０６　ＹＥＳ）、ターゲット関連性指数算出部１１４は、ターゲット関連性指数を算出

する（ステップＳ２０７）。

【０１１５】

ステップＳ２０１～ステップＳ２０７までの処理に並行して、抽出装置１００は、ステップＳ２０８～ステップＳ２１５までの処理を行う。その際、始めに抽出装置１００は、ｉ、ｊ、ｋを初期化する。次に、処理活性度算出部１２１は、ｋ番目の期間においてｉ番目のクラスタＣ＿ｉの活性度を算出する（ステップＳ２０８）。次に、活性度算出部１２１は、全てのクラスタの活性度を算出したか否か判定する（ステップＳ２０９）。活性度算出部１２１は、全てのクラスタの活性度を算出していない場合（ステップＳ２０９　ＮＯ）、ｉを１増やし（ステップＳ２１０）、ステップＳ２０８の処理に戻る。

【０１１６】

一方、活性度算出部１２１が全てのクラスタの活性度を算出した場合（ステップＳ２０９　ＹＥＳ）、活性度算出部１２１は、全期間のドキュメントで活性度を算出したか否か判定する（ステップＳ２１１）。活性度算出部１２１は、全期間のドキュメントで活性度を算出していない場合（ステップＳ２１１　ＮＯ）、ｋを１増やし（ステップＳ２１２）、ステップＳ２０８の処理に戻る。

一方、活性度算出部１２１が全期間のドキュメントで活性度を算出した場合（ステップＳ２１１　ＹＥＳ）、相対力指数算出部１２２は、ｊ番目のクラスタＣ＿ｊの相対力指数（ＲＳＩ）を算出する（ステップＳ２１３）。

【０１１７】

次に、相対力指数算出部１２２は、全てのクラスタの相対力指数（ＲＳＩ）を算出したか否か判定する（ステップＳ２１４）。相対力指数算出部１２２は、全てのクラスタの相対力指数（ＲＳＩ）を算出していない場合（ステップＳ２１４　ＮＯ）、ｊを１増やし（ステップＳ２１５）、ステップＳ２１３の処理に戻る。

一方、相対力指数算出部１２２が、全てのクラスタの相対力指数（ＲＳＩ）を算出した場合（ステップＳ２１４　ＹＥＳ）、抽出装置１００は、ステップＳ２１６の処理に進む。

【０１１８】

次に、ステップＳ２１６において、クラスタ組抽出部１３０は、活性化予測条件を満たす下で、新規発見性指数とターゲット関連性指数とに基づいた評価値が最大になるクラスタの組み合わせを抽出する（ステップＳ２１６）。以上で、本フローチャートの処理を終了する。

【０１１９】

以上により、本実施形態の抽出装置１００は、抽出された３つのクラスタのうち少なくとも１つが活性化されていること、抽出された２つのクラスタの組み合わせに意外性があること、その２つのクラスタの組み合わせは直接の関連性は薄いが、抽出されたもう１つのクラスタ（第３のクラスタ）を経由すると結び付けられるものであること、そのクラスタの組み合わせを提供する対象であるターゲットの特性と抽出されたクラスタのうち少なくとも１つとが関連性があることという条件下で、クラスタの組み合わせを提供することができる。各クラスタは１つの概念と対応しているので、抽出装置１００は、所定の期間において、そのターゲットにとって意外性があり、第３のクラスタに対応する第３の概念を介して結び付けられる概念の組み合わせを提供することができる。

【０１２０】

また、本実施形態の抽出装置１００の各処理を実行するためのプログラムをコンピュータ読み取り可能な記録媒体に記録して、当該記録媒体に記録されたプログラムをコンピュータシステムに読み込ませ、実行することにより、抽出装置１００に係る上述した種々の処理を行ってもよい。

【０１２１】

なお、ここでいう「コンピュータシステム」とは、ＯＳや周辺機器等のハードウェアを含むものであってもよい。また、「コンピュータシステム」は、ＷＷＷシステムを利用している場合であれば、ホームページ提供環境（あるいは表示環境）も含むものとする。ま

た、「コンピュータ読み取り可能な記録媒体」とは、フレキシブルディスク、光磁気ディスク、ＲＯＭ、フラッシュメモリ等の書き込み可能な不揮発性メモリ、ＣＤ－ＲＯＭ等の可搬媒体、コンピュータシステムに内蔵されるハードディスク等の記憶装置のことをいう。

【０１２２】

さらに「コンピュータ読み取り可能な記録媒体」とは、インターネット等のネットワークや電話回線等の通信回線を介してプログラムが送信された場合のサーバやクライアントとなるコンピュータシステム内部の揮発性メモリ（例えばＤＲＡＭ（Ｄｙｎａｍｉｃ Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ））のように、一定時間プログラムを保持しているものも含むものとする。また、上記プログラムは、このプログラムを記憶装置等に格納したコンピュータシステムから、伝送媒体を介して、あるいは、伝送媒体中の伝送波により他のコンピュータシステムに伝送されてもよい。ここで、プログラムを伝送する「伝送媒体」は、インターネット等のネットワーク（通信網）や電話回線等の通信回線（通信線）のように情報を伝送する機能を有する媒体のことをいう。また、上記プログラムは、前述した機能の一部を実現するためのものであっても良い。さらに、前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるもの、いわゆる差分ファイル（差分プログラム）であっても良い。

【０１２３】

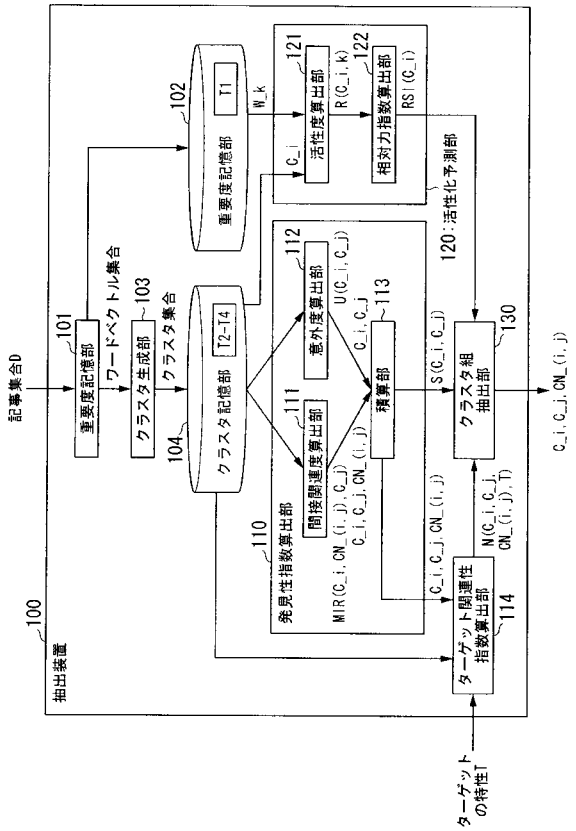
以上、本発明の実施形態について図面を参照して詳述したが、具体的な構成はこの実施形態に限られるものではなく、この発明の要旨を逸脱しない範囲の設計等も含まれる。

【符号の説明】

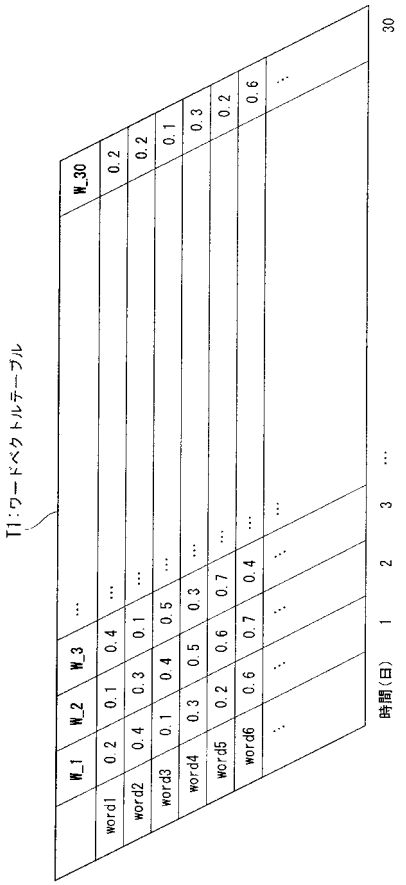
【０１２４】

１００　抽出装置
１０１　重要度算出部
１０２　重要度記憶部
１０３　クラスタ生成部
１０４　クラスタ記憶部
１１０　発見性指数算出部
１１１　間接関連度算出部
１１２　意外度算出部
１１３　積算部
１１４　ターゲット関連性指数算出部
１２０　活性化予測部
１２１　活性度算出部
１２２　相対力指数算出部（活性度上昇期待値算出部）
１３０　クラスタ組抽出部

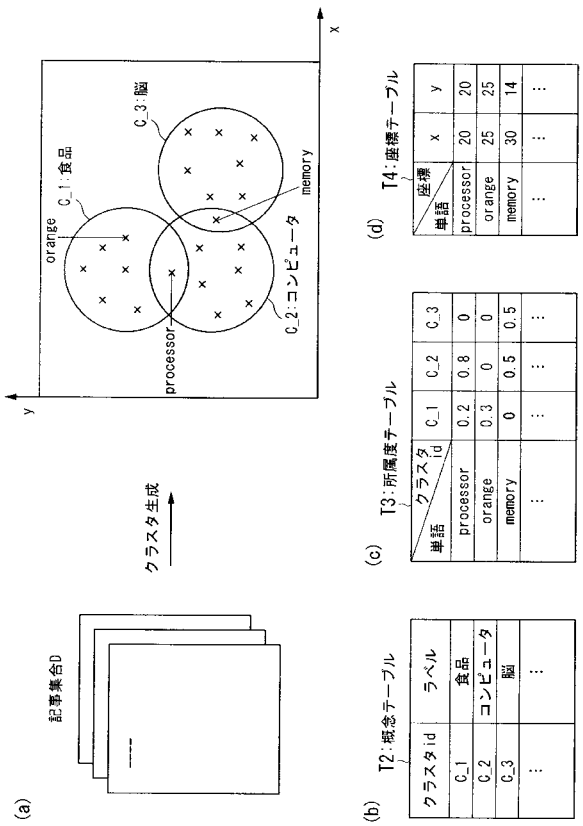
【図１】


記事集合D
抽出装置
100
重要度記憶部
101
ワードベクトル集合
クラスタ生成部
103
クラスタ集合
クラスタ記憶部
T2-T4
104
重要度記憶部
T1
102
W_k
C_i
活性度算出部
121
R(C_i,k)
相対力指数算出部
122
RS(C_i)
120:活性化予測部
意外度算出部
112
U(C_i,C_j)
C_i,C_j
113
積算部
S(C_i,C_j)
クラスタ組抽出部
130
C_i,C_j,CN_(i,j)
発見性指数算出部
110
間接関連度算出部
111
MIR(C_i,CN_(i,j),C_j)
C_i,C_j,CN_(i,j)
C_i,C_j,CN_(i,j)
ターゲット関連性指数算出部
114
N(C_i,C_j,CN_(i,j),T)
ターゲットの特性T

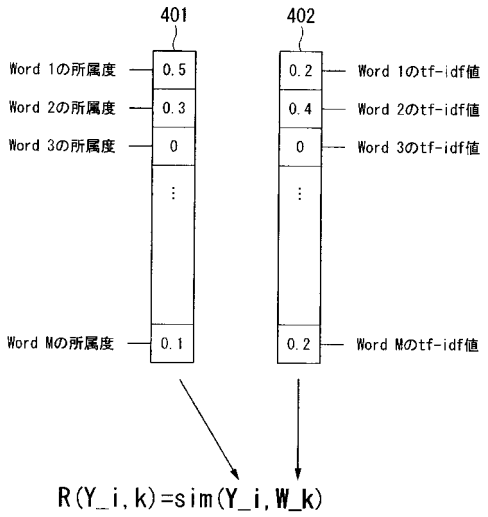
【図２】


T1:ワードベクトルテーブル
W_1
W_2
W_3
…
W_30
word1
word2
word3
word4
word5
word6
時間(日)
1
2
3
…
30

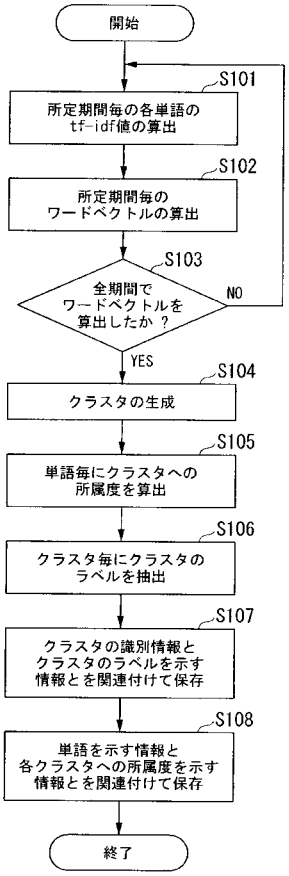
【図３】


(a)
記事集合D
クラスタ生成
(b)
T2:概念テーブル
クラスタid
ラベル
C_1
食品
C_2
コンピュータ
C_3
脳
(c)
T3:所属度テーブル
単語
クラスタid
processor
orange
memory
(d)
T4:座標テーブル
座標
x
y
C_1:食品
C_2:コンピュータ
C_3:脳
orange
processor
memory

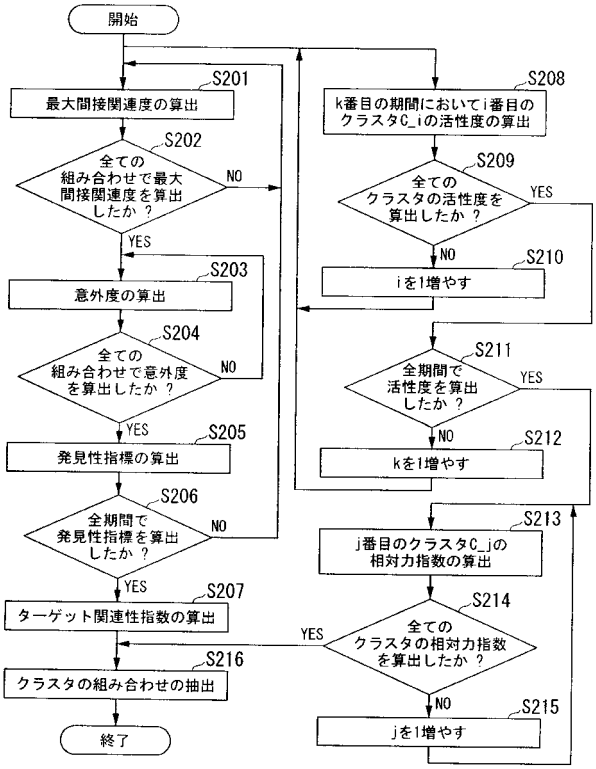
【図４】


401
402
Word 1の所属度
Word 2の所属度
Word 3の所属度
Word Mの所属度
Word 1のtf-idf値
Word 2のtf-idf値
Word 3のtf-idf値
Word Mのtf-idf値
R(Y_i,k)=sim(Y_i,W_k)

【図５】


開始
S101
所定期間毎の各単語の
tf-idf値の算出
S102
所定期間毎の
ワードベクトルの算出
S103
全期間で
ワードベクトルを
算出したか ？
NO
YES
S104
クラスタの生成
S105
単語毎にクラスタへの
所属度を算出
S106
クラスタ毎にクラスタの
ラベルを抽出
S107
クラスタの識別情報と
クラスタのラベルを示す
情報とを関連付けて保存
S108
単語を示す情報と
各クラスタへの所属度を示す
情報とを関連付けて保存
終了

【図６】


開始
S201
最大間接関連度の算出
S202
全ての
組み合わせで最大
間接関連度を算出
したか ？
NO
YES
S203
意外度の算出
S204
全ての
組み合わせで意外度
を算出したか ？
NO
YES
S205
発見性指標の算出
S206
全期間で
発見性指標を算出
したか ？
NO
YES
S207
ターゲット関連性指数の算出
S216
クラスタの組み合わせの抽出
終了
S208
k番目の期間においてi番目の
クラスタC_iの活性度の算出
S209
全ての
クラスタの活性度を
算出したか ？
YES
NO
S210
iを1増やす
S211
全期間で
活性度を算出
したか ？
YES
NO
S212
kを1増やす
S213
j番目のクラスタC_jの
相対力指数の算出
S214
全ての
クラスタの相対力指数
を算出したか ？
YES
NO
S215
jを1増やす