(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2004-178123
(P2004-178123A)
(43)公開日　平成16年6月24日(2004.6.24)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/30	G06F 17/30	210D	5B075
	G06F 17/30	170A	
	G06F 17/30	180A	

審査請求 未請求　請求項の数 10　OL　（全 17 頁）

(21)出願番号　特願2002-341671(P2002-341671)
(22)出願日　平成14年11月26日(2002.11.26)

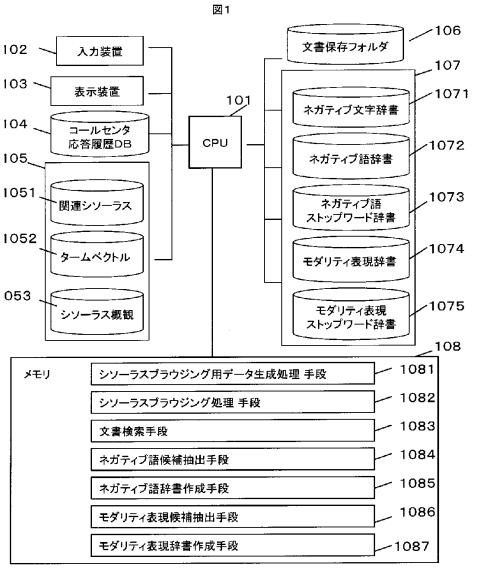
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(74)代理人　100075096
弁理士　作田　康夫
(72)発明者　小泉　敦子
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所中央研究所内
(72)発明者　森本　康嗣
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所中央研究所内
(72)発明者　隈井　裕之
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所中央研究所内

最終頁に続く

(54)【発明の名称】情報処理装置、該情報処理装置を実現するためのプログラム

(57)【要約】
【課題】従来のキーワードによる文書分類技術は、高頻度知識の抽出・分類に適しているが、コールセンターの応答履歴からリスク管理上有用な情報や顧客の生の声を抽出するには、大量のありふれた情報を取り除いた中から真に有用な知識を抽出する必要がある。
【解決手段】キーワードで検索した文書をフォルダに保存する機能を設け、キーワード検索により高頻度情報をフォルダに保存した後、残りの文書を低頻度情報のフォルダに保存する機能を設ける。低頻度情報からリスク管理上有用な知識を抽出する手段として、ネガティブな表現や心的態度を表すモダリティ表現を抽出する機能を設ける。
【選択図】　図１

図1



【特許請求の範囲】
【請求項１】
複数のデータを格納する記憶手段と、
上記記憶されるデータのうち単語若しくは語を共通に有するデータに共通の属性を付する手段と、
上記データを分析する解析手段とを有し、
上記解析手段は、属性付けのなされていないデータにはネガティブ語辞書を用いた分析を行い、上記属性付けのなされているデータには異なる分析を行うことを特徴とする情報処理装置。
【請求項２】
上記情報処理装置は、
入力手段と、
上記入力手段を介して受けつけたキーワードを用いて上記データベース内を検索する手段を有し、
上記属性を付する手段は、上記検索の結果抽出されたデータにその旨の属性付けを行うことを特徴とする請求項１記載の情報処理装置。
【請求項３】
上記入力手段は上記検索手段において抽出された回数の指定を受け、
上記解析手段は、上記回数以下抽出された旨の属性を有するデータと、上記回数より多い回数抽出された旨の属性を有するデータとで異なる解析方法で分析を行うことを特徴とする請求項２記載の情報処理装置。
【請求項４】
上記ネガティブ語辞書は、漢字単位の語を格納する第１の辞書と該漢字を含む単語を格納する第２の辞書とから構成され、
上記解析手段は、上記データから上記第１及び第２の辞書に格納される語を検索し、上記第１の辞書に格納される漢字を含むとして検索された単語のうち上記第２の辞書にないものを上記表示手段に表示して、該表示した単語のうち指定された単語を上記第２の辞書に格納することを特徴とする情報処理装置。
【請求項５】
モダリティを表現する単語を格納する辞書をさらに有し、
上記解析手段は、上記辞書を使った分析を行うことを特徴とする請求項１乃至４の何れかに記載の情報処理装置。
【請求項６】
上記記憶されるデータから単語と単語の関連度を計算する手段と、
上記記憶されるデータから重要タームを抽出する手段と、
上記関連度の情報を用いて上記重要タームをクラスタリングしシソーラス概観を生成する手段と、
上記生成されたシソーラス概観を表示手段に表示する手段とを有し
上記表示手段は、上記入力手段を介して選択された上記シソーラス概観のクラスタに属する重要タームを表示し、
上記表示される重要タームのうち上記指示入力手段を介して指示された重要タームを上記キーワードとして設定することを特徴とする請求項２乃至５の何れかに記載の情報処理装置。
【請求項７】
漢字単位の語を格納する第１の辞書と、
該漢字を含む単語を格納する第２の辞書と、
表示手段と、
入力手段と、
記録手段に記録されるデータから上記第２の辞書に格納される単語を検索する手段を有し、

上記検索手段は、上記第１の辞書に格納される漢字を含む単語も検索し、上記第１の辞書に格納される漢字を含むとして検索された単語を上記表示手段に表示して、該表示した単語のうち指定された単語を上記第２の辞書に格納することを特徴とする情報処理装置。
【請求項８】
上記指定されなかった単語を蓄積する第３の辞書を有することを特徴とする請求項７記載の情報処理装置。
【請求項９】
上記第１の辞書は否定的な意味を持つ漢字を格納し、
上記第２の辞書は、否定的な意味をもつ単語を格納することを特徴とする請求項７又は８に記載の情報処理装置。
【請求項１０】
キーワードの入力を受けるステップと、
複数のデータを格納する記憶手段に格納される複数のデータを上記キーワードを用いて検索するステップと、
上記検索の結果抽出されたデータに共通の属性を付するステップと、
上記属性付けのされていないデータについてネガティブ語辞書を用いた分析を行い、上記属性付けのなされているデータには上記ネガティブ辞書とは異なるデータを用いた分析を行うステップとをコンピュータに実行させることを特徴とするプログラム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、自然言語で記述されたテキストから知識を抽出するテキスト分析方法に関する。主として、コールセンターの応答履歴の分析を対象とする。
【０００２】
【従来の技術】
ユーザが指定したキーワードにより文書を分類する文書分類システムとしては、文書中の単語の出現頻度に基づいて未使用視点（まだ分類に使っていないキーワード）を検出し表示することによりキーワードによる分類を支援する文書分類システムがある（例えば、特許文献１参照）。
リスク管理の上で有用な知識を抽出する手段としては、「失礼」「失望」などのネガティブな表現に着目することが考えられる。ネガティブ表現を抽出する方法としては、ドメインに応じて「失注」、「苦情」などのネガティブな意味を持つキーワードを予めセットしておき、検索を実行して、ヒットした場合にはアラートを出すという方法が考えられる。更に、文書分類のためのキーワード辞書をユーザが更新する手段を設けた文書分類システムもある（例えば、特許文献２参照）。
【特許文献１】特開２００１－１０１２２６号公報
【特許文献２】特開２００１－１８４３５１号公報
【発明が解決しようとする課題】
従来のキーワードによる文書分類技術は、高頻度知識の抽出・分類に適しているが、コールセンターの応答履歴からリスク管理上有用な情報や顧客の生の声を抽出するには、低頻度の知識の抽出が重要課題である。すなわち、大量のありふれた情報を取り除いた中から、効率よく、かつ漏れなく、真に有用な知識を抽出する必要がある。本発明の目的は、高頻度の問合せに基づいてＦＡＱを作成することと、低頻度の問合せの中からリスク管理上有用な情報を抽出することにある。
リスク管理の目的でテキスト分析を行う際に、ネガティブな表現を抽出することが考えられる。ネガティブな表現を抽出するためには、ドメインに応じて「失望」「失礼」などのキーワードをセットしておき、検索を実行する方法が考えられるが、予めキーワードを設定することに手数がかかる上に、網羅することが困難であり、漏れが多く発生するという問題がある。
【０００３】

【課題を解決するための手段】
上記課題を解決するため、テキスト分析支援システムにおいて、低頻度情報を抽出するための手段として、高頻度情報を含む文書を抽出してフォルダに保存した後、残りの文書を集めて低頻度情報のフォルダに保存する機能を設け、低頻度情報のフォルダのデータにはネガティブ表現の抽出漏れとノイズをなくすための手段として、「失」「負」などのネガティブな意味を持つ文字を格納した辞書を用いて対象テキストからネガティブ語候補を抽出し、ネガティブ語と判定したものをネガティブ語辞書に登録した上で、ネガティブ語辞書を用いてネガティブ表現の抽出を行うようにする。
また、
【０００４】
【発明の実施の形態】
以下、本発明の実施例について説明する。本実施例は、コールセンタの応答履歴を対象としたテキスト分析支援システムである。以下、図面を使って詳細に説明する。
（システム構成）
図１は本発明の第１の実施例を示すテキスト分析支援システムの構成図である。本システムは、ＣＰＵ１０１、入力装置１０２、表示装置１０３、コールセンタ応答履歴データベース１０４、シソーラスブラウジング用データ格納部１０５、文書保存フォルダ１０６、低頻度知識抽出用データ格納部１０７、メモリ１０８によって構成されている。シソーラスブラウジング用データ格納部１０５は、関連シソーラス格納部１０５１、タームベクトル格納部１０５２、およびシソーラス概観格納部１０５３によって構成されている。低頻度知識抽出用データ格納部１０７は、ネガティブ表現抽出機能を実現するためのネガティブ文字辞書１０７１、ネガティブ語辞書１０７２、ネガティブ語ストップワード辞書１０７３、モダリティ表現抽出機能を実現するためのモダリティ表現辞書１０７４、モダリティ表現ストップワード辞書１０７５によって構成されている。メモリ１０８には、シソーラスブラウジング用データ生成処理手段１０８１、シソーラスブラウジング処理手段１０８２、文書検索手段１０８３、ネガティブ語候補抽出手段１０８４、ネガティブ語辞書作成手段１０８５、モダリティ表現候補抽出手段１０８６、モダリティ表現辞書作成手段１０８７が記憶されている。
（コールセンタ応答履歴データベース）
図２にコールセンタ応答履歴データベース１０４のデータ構造を示す。コールセンタ応答履歴データベース１０４の各レコードには、問合せＩＤ１０４１、応答履歴メモ１０４２、キーワード検索で検索済みであることを示す検索フラグ１０４３、分類フォルダに分類済みであることを示す分類フラグ１０４４が記述されている。
（シソーラスブラウジング機能）
本システムは、高頻度情報を含む文書の抽出を支援するシソーラスブラウジング機能を備えている。ここでいうシソーラスとは、文書群中の特徴的な単語とその関係を示すネットワーク表現である。本システムのシソーラスブラウジング機能は、文書群からシソーラスを自動生成する機能と、生成したシソーラスの概観や細部を表示する機能（概観表示・ズーム表示）からなる。シソーラス自動生成およびシソーラス表示は、例えば特開２０００－２２７９１７に記載されているシソーラスブラウジング方法によって行う。以下、本システムにおいてシソーラスブラウジング機能を実現するためのデータおよび処理手順の概要を説明する。まず、シソーラスブラウジング機能を実現するためのデータについて説明する。シソーラスブラウジング用データ格納部１０５は、関連シソーラス格納部１０５１、タームベクトル格納部１０５２、およびシソーラス概観格納部１０５３によって構成されている。
関連シソーラス格納部１０５１には、コールセンタ応答履歴データベース１０４の応答履歴メモ１０４２に格納された文書データから生成した関連シソーラスが格納されている。関連シソーラスとは、単語と単語の関連度を示すものである。本実施例では、関連度は２つの単語の共起しやすさを表すものであり、それぞれの単語の頻度と共起頻度（文書中のある範囲内に２つの語が同時に出現する頻度）に基づいて計算される。図３に関連シソー

ラス格納部１０５１のデータ構造を示す。関連シソーラス格納部１０５１は、レコードＩＤ１０５１１、タームＸ１０５１２、タームＹ１０５１３、および関連度１０５１４から構成される。タームＸ１０５１２およびタームＹ１０５１３には、関連関係にあるタームを、関連度１０５１４にはその関連度を格納する。
タームベクトル格納部１０５２には、コールセンタ応答履歴データベース１０４の応答履歴メモ１０４２に格納された文書データから抽出したタームベクトルが格納されている。タームベクトルとは、文書を特徴付けるタームのリストであり、「Ｓａｌｔｏｎ，Ｇ．，ｅｔａｌ．：Ａ Ｖｅｃｔｏｒ Ｓｐａｃｅ Ｍｏｄｅｌ ｆｏｒ Ａｕｔｏｍａｔｉｃ Ｉｎｄｅｘｉｎｇ，Ｃｏｍｍｕｎｉｃａｔｉｏｎｓ ｏｆ ｔｈｅ ＡＣＭ，Ｖｏｌ．１８，Ｎｏ．１１（１９７５）．」に記載のｔｆ－ｉｄｆ法（Ｔｅｒｍ Ｆｒｅｑｕｅｎｃｙ ｉｎｖｅｒｓｅ Ｄｏｃｕｍｅｎｔ Ｆｒｅｑｕｅｎｃｙ）を利用することにより抽出可能である。このｔｆ－ｉｄｆ法は、文書インデクシング方法として最もよく知られているもののひとつであり、ある文書におけるタームの出現頻度（ｔｆ）と、当該タームが出現した文書数の逆数（ｉｄｆ）をかけた値を当該文書におけるタームの重要度とし、当該文書において重要度の高いターム（すなわち重要ターム）を抽出してタームベクトルとする技術である。図４にタームベクトル格納部１０５２のデータ構造を示す。タームベクトル格納部１０５２は、レコードＩＤ１０５２１、問合せＩＤ１０５２２および重要タームリスト１０５２３から構成される。問合せＩＤ１０５２１には、コールセンタ応答履歴データベース１０４に格納された応答履歴のＩＤを格納し、重要タームリスト１０５２２には当該応答履歴の応答メモに出現するタームのうち重要なもののリストが格納される。
シソーラス概観格納部１０５３には、関連シソーラス格納部１０５１に格納された関連シソーラスの概観が格納されている。シソーラス概観とは、文書群中のもっとも特徴的な単語を代表タームとして抽出し、関係の強い代表タームをタームクラスタとしてまとめたものである。図５にシソーラス概観格納部１０５３のデータ構造を示す。シソーラス概観格納部１０５３は、タームグループ番号１０５３１およびタームリスト１０５３２から構成される。タームリスト１０５３２には、タームクラスタに属するタームのリストが格納される。

【０００５】

以上、シソーラスブラウジング用データについて説明した。
次に、シソーラスブラウジング機能を実現するためのシソーラスブラウジング用データ生成処理手順および、シソーラスブラウジング処理手順について図７および図８のフローチャートを用いて説明する。
（シソーラスブラウジング用データ生成処理手順）
まず、分析環境準備として、シソーラスブラウジング用データを作成する。図７に示すように、シソーラスブラウジング用データ生成処理では、まず文書データからタームとタームの関連度を示す関連シソーラスを生成し（ステップ７０１）、各文書のタームベクトルを抽出して（ステップ７０２）、シソーラス概観を生成する（ステップ７０３）。シソーラス概観は、文書群中のもっとも特徴的な単語を代表タームとして抽出し、関係の強い代表タームをタームクラスタとしてまとめたものである。代表ターム抽出処理では、各文書タームベクトルを構成する重要タームのうち、多くの文書で重要タームとなったタームを代表タームとする。タームクラスタ生成処理では、関連シソーラスに格納されたターム間の関連度に基づいて関連度の高い代表タームをひとつのクラスタにまとめる。
（シソーラスブラウジング処理手順）
図８に示すように、シソーラスブラウジング処理では、まずシソーラス概観格納部１０５３に格納されたシソーラス概観を例えば図６のシソーラス概観表示部６０２に示すような形でユーザに表示する（ステップ８０１）。シソーラス概観表示部６０２は、タームリスト表示部６０２１および選択ボタン６０２２からなる。タームリスト表示部６０２１には、シソーラス概観格納部１０５３に格納されているタームリスト１０５３２が表示される。次にユーザがタームクラスタリスト６０２１を選択ボタン等の指示入力手段６０２２で

選択してズームボタン６０３３でズームを指示すれば（ステップ８０２）、ユーザが選択したタームクラスタに属するタームの関連タームを関連シソーラス１０５１より取得する（ステップ８０３）。そして、それらをクラスタリングし（ステップ８０４）、生成したタームクラスタを関連タームクラスタ表示部６０４に表示する（ステップ８０５）。ユーザからのシソーラスブラウジング終了の指示があれば（ステップ８０６）、処理を終了し、なければステップ８０２の処理に戻る。ステップ８０２のズーミング指示において、関連タームクラスタ表示部６０４に表示されているタームクラスタ６０４１を選択ボタン６０４２で選択してズームボタン６０３３でズームを指示すれば、該関連タームクラスタの関連語が関連タームクラスタ表示部６０４に表示される。また、シソーラス概観表示部６０２あるいはタームクラスタ表示部６０４に表示されているタームをクリックしてからズームボタン６０３３をクリックすると、該タームの関連語が関連タームクラスタ表示部６０４に表示される。ユーザは、関連クラスタ数６０３１およびクラスタ内ターム数６０３３を選択することにより、いくつのクラスタに分けるか、１つのクラスタについて何ターム抽出するかを指定することができる。

（シソーラスブラウジングによる効果）

このようにキーワードで文書を検索する機能と、検索した文書をフォルダに保存する機能を設け、ユーザがキーワードとして入力した語に関連する問合せを抽出し、ＦＡＱ作成のために保存することができるようにする。また、応答履歴全体からシソーラスを生成し、シソーラスの全体構造を示すシソーラス概観から、ユーザが選択したタームを含む部分構造へと、ユーザをナビゲートするシソーラスブラウジング機能を設け、ユーザがキーワードを想起しやすいようにする。シソーラス概観を眺めることにより、文書群中のトピックを俯瞰することができる。１つのタームクラスタにまとめられた代表タームの並びを見ると、トピックやその内容を推測することができる。タームの関連語をクラスタ表示（関係の強い語をタームクラスタとしてまとめて表示）することにより、タームに対応するトピックのサブトピックとその内容を推測することができる。

【０００６】

本システムは、シソーラスブラウジング機能およびキーワード文書検索機能により高頻度情報を含む文書を抽出して分類フォルダに保存した後、残りの文書を集めて低頻度情報のフォルダに保存する機能を備えている。図６に文書分類操作画面の構成を示す。図６に示すように、文書分類操作画面６０１は、シソーラスブラウジング機能のためのシソーラス概観表示部６０２、シソーラスズーミング指示部６０３、関連タームクラスタ表示部６０４、キーワード文書検索機能のための文書検索指示部６０５、文書検索結果表示部６０６、文書分類保存機能のための文書保存部６０７からなる。

シソーラス概観表示部６０２は、タームリスト表示部６０２１および選択ボタン６０２２からなる。タームリスト表示部６０２１には、シソーラス概観格納部１０５３に格納されているタームリスト１０５３２が表示される。シソーラスズーミング指示部６０３は、クラスタ数６０３１、クラスタ内ターム数６０３２、ズームボタン６０３３からなる。

関連タームクラスタ表示部６０４は、タームリスト表示部６０４１および選択ボタン６０４２からなる。

文書検索指示部６０５は、検索ターム入力部６０５１および検索ボタン６０５２からなる。文書検索結果表示部６０６は、文書表示部６０６１および文書選択ボタン６０６２からなる。文書保存部６０７はフォルダ名表示部６０７１およびフォルダ選択ボタン６０７２からなる。

（文書分類手順）

本システムは、高頻度情報を含む文書を抽出してフォルダに保存した後、残りの文書を集めて低頻度情報のフォルダに保存する機能を備えている。図９は、本システムによる文書分類手順を示すフローチャートである。本システムによる文書分類手順について、図６の文書分類操作画面および図９のフローチャートを用いて説明する。まず、分類開始指示があると（ステップ９０１）、コールセンタ応答履歴データベース１０４にアクセスし、検索済みであることを示す検索フラグ１０４３と、分類済みであることを示す分類フラグ１

０４４の値を“０”にリセットする。ユーザがターム入力部６０５１にタームを入力し、検索ボタン６０５２をクリックしてキーワード文書検索を指示すると（ステップ９０３）、コールセンタ応答履歴データベース１０４の応答履歴メモ１０４２を対象にキーワード文書検索を行い（ステップ９０４）、コールセンタ応答履歴データベース１０４の検索フラグ１０４３に検索済みであることを示すフラグ“１”を設定し（ステップ９０５）、文書検索結果を文書検索結果表示部６０６の文書表示部６０６１に表示する（ステップ９０６）。ユーザが文書検索結果一覧から保存したい文書を選択して文書選択ボタン６０６２とフォルダ選択ボタン６０７２をクリックすると（ステップ９０７）、選択された文書を文書保存フォルダ１０６へ保存し（ステップ９０８）、コールセンタ応答履歴データベース１０４の分類フラグ１０４４に分類済みであることを示すフラグ“１”を設定する（ステップ９０９）。ユーザから分類終了の指示があれば（ステップ９１０）、検索済みフラグ＝０の文書を低頻度文書フォルダに保存する（９１１）。

低頻度文書フォルダへの文書保存方法の代案としては、分類済みフラグ＝０の文書を低頻度文書フォルダに保存するようにしてもよい。また、文書保存フォルダに選択フラグを用意し、ユーザが指定したフォルダに分類済みの文書以外の文書を低頻度文書フォルダに保存するようにしてもよい。さらに、検索済み、分類済みかどうかを示す検索フラグおよび分類済みフラグの変わりに検索回数および分類回数を更新するようにし、検索回数あるいは分類回数が閾値よりも低いものを低頻度文書フォルダに保存するようにしてもよい。

【０００７】

本システムは、キーワード想起を支援するシソーラスブラウジング機能を備えている。ユーザは、シソーラスブラウジングの過程で、表示されたタームを選択することによりキーワード文書検索を行うこともできる。シソーラス概観表示部６０２のタームリスト表示部６０２１に表示されたタームをクリックすると該タームが検索ターム入力部６０５１にコピーされる。また、シソーラス概観表示部６０２の選択ボタン６０２２をクリックすると、タームリスト表示部６０２１に表示されている全てのタームが検索ターム入力部６０５１にコピーされる。同様に、関連タームクラスタ表示部６０４のタームリスト表示部６０４１に表示されたタームをクリックすると該タームが検索ターム入力部６０５１にコピーされ、選択ボタン６０４２をクリックすると、タームリスト表示部６０４１に表示されている全てのタームが検索ターム入力部６０５１にコピーされる。シソーラスには、応答履歴全体に出現するタームが関連付けて格納されている。したがって、シソーラスブラウジングをすることにより、高頻度情報を収集・分類することができる。

（低頻度情報からの知識抽出）

以上に述べたように、本システムでは、分類開始から終了までの間に一度も検索されていな文書、あるいは、どの分類フォルダにも分類されていない文書をまとめて低頻度情報フォルダに格納することができる。リスク管理の目的でテキスト分析を行う際に、失礼」「失望」などのネガティブな意味を持つ単語や、「くれないのか」「そもそも」「なんなのか」「欲しい」などのモダリティ表現が有効な手がかりとなる。そこで、低頻度情報からリスク管理上有用な知識を抽出する手段として、ネガティブな表現を抽出する機能と、顧客やオペレータの心的態度を表すモダリティ表現を抽出する機能を設ける。以下、低頻度情報フォルダに保存された応答履歴メモからネガティブ表現およびモダリティ表現を含む文書を抽出する手順の概要を図２１のフローチャートに従って説明する。まず、低頻度情報フォルダに保存された応答履歴メモから、ネガティブ語候補・モダリティ表現候補を抽出する（ステップ２１０１）。次に、ネガティブ語候補・モダリティ表現候補のうち、ユーザが選択したものをネガティブ語辞書・モダリティ表現辞書に登録する（ステップ２１０２）。最後に、低頻度情報フォルダの文書に対して、ネガティブ語辞書およびモダリティ表現辞書に登録された語をキーワードとしてキーワード検索を行うことにより（ステップ２１０３）、ネガティブ語およびモダリティ表現を含む文書を抽出し、内容を確認する（ステップ２１０４）。

以下、ネガティブ表現およびモダリティ表現の抽出の手順について詳細に述べる。

（ネガティブ表現の抽出）

応答履歴メモからネガティブな表現を抽出する手段として、本システムは、応答履歴メモからネガティブ語候補を抽出するネガティブ語候補抽出機能と、ネガティブ語候補の中でユーザがネガティブ語と判定した語をネガティブ語辞書に登録するネガティブ語辞書作成機能とを備えている。これらの機能を実現するため、本システムは、「失」「負」「遅」などのネガティブ語の構成要素となりやすい文字を登録したネガティブ文字辞書１０７１、ネガティブ語であることが判定済みの語が登録されているネガティブ語辞書１０７２、ネガティブ語でないことが判定済みの語が登録されているネガティブ語ストップワード辞書１０７３を備えている。

図１２に、ネガティブ文字辞書１０７１のデータ構造を示す。ネガティブ文字辞書の各レコードには、レコードＩＤ１０７１１、ネガティブ文字１０７１２、ネガティブ度１０７１３、ネガティブ語辞書登録語数１０７１４、ネガティブ語ストップワード辞書登録語数１０７１５が記述されている。ネガティブ語辞書登録語数１０７１４は、ネガティブ語辞書に登録されている単語のうち、当該ネガティブ文字を含む単語の語数である。ネガティブ語ストップワード辞書登録語数１０７１５は、ネガティブ語ストップワード辞書１０７３に登録されている単語のうち、当該ネガティブ文字を含む単語の語数である。ネガティブ度１０７１３には、ネガティブ語候補として抽出された単語のうちネガティブ語辞書に登録された単語の割合を示す０～１の値が記述されている。あるいは、ネガティブ度の値はユーザが任意に設定するようにしてもよい。図１３に、ネガティブ語辞書１０７２のデータ構造を示す。ネガティブ語辞書の各レコードには、レコードＩＤ１０７２１、ネガティブ語１０７２２、ネガティブ度１０７２３が記述されている。ネガティブ度１０７２３には、ネガティブ文字辞書に記述されたネガティブ度１０７１３の値が記述されている。

図１４に、ネガティブ語ストップワード辞書１０７３のデータ構造を示す。ネガティブ語ストップワード辞書の各レコードには、レコードＩＤ１０７３１、ネガティブ語ストップワード１０７３２が記述されている。

以下、ネガティブ語候補抽出の手順を図１７のフローチャートにしたがって説明する。まず、応答履歴メモ１０４２にあらわれるすべての単語を抽出し、単語リストを作成する（ステップ１７０１）。単語リストの単語を１語読み（ステップ１７０３）、ネガティブ文字辞書１０７１を参照し、ネガティブ文字を含むかどうかを判定する（ステップ１７０４）。ネガティブ文字を含む場合は、ネガティブ語辞書１０７２を参照し、ネガティブ語辞書１０７２に登録済みであるかどうかを判定する（ステップ１７０５）。ネガティブ語辞書１０７２に登録済みの場合は、ネガティブ語であることがすでにわかっているので、ネガティブ語候補として抽出せずにこの単語に関する処理を終了する。ネガティブ語辞書１０７２に未登録の場合は、ネガティブ語ストップワード辞書１７０３を参照し、ネガティブ語ストップワード辞書１０７３に登録済みであるかどうかを判定する（ステップ１７０６）。ネガティブ語ストップワード辞書１０７３に登録済みの場合は、ネガティブ語でないことがすでにわかっているので、ネガティブ語候補として抽出せずにこの単語に関する処理を終了する。そして、ネガティブ語辞書にもネガティブ語ストップワード辞書にも登録されていない単語をネガティブ語候補リストに登録する（ステップ１７０７）。単語リストに登録されているすべての単語について同様の処理を行うことにより、ネガティブ文字を含む単語のうち、ネガティブ語辞書にもネガティブ語ストップワード辞書にも登録されていない単語をネガティブ語候補リストに登録する。

以下、ネガティブ語辞書作成の手順を図１８のフローチャートにしたがって説明する。まず、ネガティブ語候補に対してネガティブ語かどうかの判定を行うため、ネガティブ語候補リストを画面に表示する（ステップ１８０１）。図１１にネガティブ語判定画面の表示例を示す。ネガティブ語判定画面には、ネガティブ語候補表示部１１０１１、ネガティブ語辞書既登録語表示部１１０１２、ネガティブ語ストップワード辞書既登録語表示部１１０１３、登録ボタン１１０１４が配置されている。ネガティブ語辞書既登録語表示部１１０１２およびネガティブ語ストップワード辞書既登録語表示部１１０１３は判定のための参考情報として表示するものだが、省いても良い。ユーザは、ネガティブ語候補表示部１１０１１に表示されたネガティブ語候補に対してネガティブ語かどうかを判定し、ネガテ

ィブ語と判定した語にチェックマークをいれる（ステップ１８０２）。ユーザが登録ボタン１１０１４をクリックすると（ステップ１８０３）、ネガティブ語と判断された語がネガティブ語辞書に登録される（ステップ１８０４）。ネガティブ語と判断されなかった語は、ネガティブ語ストップワード辞書に登録される（ステップ１８０５）。

（モダリティ表現の抽出）

次に、顧客やオペレータの心的態度を表すモダリティ表現を抽出する機能について述べる。図１５に、モダリティ表現辞書１０７４のデータ構造を示す。モダリティ表現辞書の各レコードには、レコードＩＤ１０７４１、モダリティ表現１０７４２、品詞１０７４３、モダリティ１０７４４が記述されている。図１６に、モダリティ表現ストップワード辞書１０７５のデータ構造を示す。モダリティ表現ストップワード辞書の各レコードには、レコードＩＤ１０７５１、モダリティ表現ストップワード１０７５２、品詞１０７５３が記述されている。

【０００８】

以下、モダリティ表現候補抽出の手順を図１９のフローチャートにしたがって説明する。まず、応答履歴メモ１０４２にあらわれるすべての単語を抽出し、単語リストを作成する（ステップ１９０１）。単語リストの単語を１語読み（ステップ１９０３）、品詞が副詞か助動詞の場合は（ステップ１９０４）、モダリティ表現候補抽出の処理を進める。すなわち、モダリティ表現辞書１０７４を参照し、モダリティ表現辞書１０７４に登録済みであるかどうかを判定する（ステップ１９０５）。モダリティ表現辞書１０７４に登録済みの場合は、モダリティ表現であることがすでにわかっているので、モダリティ表現候補として抽出せずにこの単語に関する処理を終了する。モダリティ表現辞書１０７４に未登録の場合は、モダリティ表現ストップワード辞書１７０５を参照し、モダリティ表現ストップワード辞書１０７５に登録済みであるかどうかを判定する（ステップ１９０６）。モダリティ表現ストップワード辞書１０７５に登録済みの場合は、モダリティ表現でないことがすでにわかっているので、モダリティ表現候補として抽出せずにこの単語に関する処理を終了する。そして、モダリティ表現辞書にもモダリティ表現ストップワード辞書にも登録されていない単語をモダリティ表現候補リストに登録する（ステップ１９０７）。単語リストに登録されているすべての単語について同様の処理を行うことにより、品詞が副詞あるいは助動詞である単語のうち、モダリティ表現辞書にもモダリティ表現ストップワード辞書にも登録されていない単語をモダリティ表現候補リストに登録する。

以下、モダリティ表現辞書作成の手順を図２０のフローチャートにしたがって説明する。まず、モダリティ表現候補に対してモダリティ表現かどうかの判定を行うため、モダリティ表現候補リストを画面に表示する（ステップ２００１）。モダリティ表現判定画面は、図１１のネガティブ語判定画面と同様のものを用いる。ユーザは、画面に表示されたモダリティ表現候補に対してモダリティ表現かどうかを判定し、モダリティ表現と判定した語にチェックマークをいれる（ステップ２００２）。ユーザが登録ボタンをクリックすると（ステップ２００３）、モダリティ表現と判断された語がモダリティ表現辞書に登録される（ステップ２００４）。モダリティ表現と判断されなかった語は、モダリティ表現ストップワード辞書に登録される（ステップ１８０５）。

【０００９】

【発明の効果】

本発明によれば、応答履歴メモに含まれる情報を高頻度情報と低頻度情報に分けることができ、それぞれに適したテキスト分析方法を適用することができるという効果がある。高頻度情報に対しては、トピックで分類することにより、ＦＡＱ作成支援に活用することができる。低頻度情報に対しては、ネガティブ表現およびモダリティ表現というトピックとは別の観点から、リスク管理上有用な知識を抽出することができる。

本発明のネガティブ表現抽出方法によれば、文字を手がかりにして分析対象テキストに含まれるネガティブ語候補を抽出するので、抽出漏れを防ぐことができる。抽出したネガティブ語候補についてネガティブ語かどうかの判定を人手で行う必要があるが、ネガティブ語かどうか判定済みの語をネガティブ語辞書およびネガティブ語ストップワード辞書に蓄

積していくので、繰り返すうちにネガティブ語候補として抽出されるものが減っていくという効果がある。

【図面の簡単な説明】

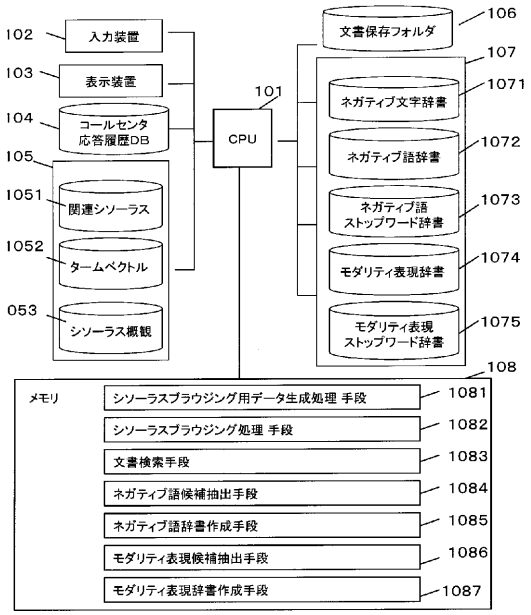
【図１】本発明のテキスト分析支援システムの実施例のシステム構成図である。
【図２】コールセンター応答履歴データベースのデータ構造を示す図である。
【図３】関連シソーラス格納部のデータ構造を示す図である。
【図４】タームベクトル格納部のデータ構造を示す図である。
【図５】シソーラス概観格納部のデータ構造を示す図である。
【図６】文書分類操作画面の構成を示す図である。
【図７】シソーラスブラウジング用データ生成処理手順を示すフローチャートである。
【図８】シソーラスブラウジング処理手順を示すフローチャートである。
【図９】文書分類手順を示すフローチャートである。
【図１０】文書保存フォルダのデータ構造を示す図である。
【図１１】ネガティブ語判定画面の表示例を示す図である。
【図１２】ネガティブ文字辞書のデータ構造を示す図である。
【図１３】ネガティブ語辞書のデータ構造を示す図である。
【図１４】ネガティブ語ストップワード辞書のデータ構造を示す図である。
【図１５】モダリティ表現辞書のデータ構造を示す図である。
【図１６】モダリティ表現ストップワード辞書のデータ構造を示す図である。
【図１７】ネガティブ語候補抽出手順を示すフローチャートである。
【図１８】ネガティブ語辞書作成手順を示すフローチャートである。
【図１９】モダリティ表現候補抽出手順を示すフローチャートである。
【図２０】モダリティ表現辞書作成手順を示すフローチャートである。
【図２１】ネガティブ表現およびモダリティ表現の抽出手順を示すフローチャートである。

【符号の説明】

１０１：ＣＰＵ
１０２：入力装置
１０３：表示装置
１０４：コールセンタ応答履歴データベース
１０５：シソーラスブラウジング用データ格納部
１０６：文書保存フォルダ
１０７：低頻度知識抽出用データ格納部
１０８：メモリ
１０５１：関連シソーラス格納部
１０５２：タームベクトル格納部
１０５３：およびシソーラス概観格納部
１０７１：ネガティブ文字辞書
１０７２：ネガティブ語辞書
１０７３：ネガティブ語ストップワード辞書
１０７４：モダリティ表現辞書
１０７５：モダリティ表現ストップワード辞書
１０８１：シソーラスブラウジング用データ生成処理手段
１０８２：シソーラスブラウジング処理手段
１０８３：文書検索手段
１０８４：ネガティブ語候補抽出手段
１０８５：ネガティブ語辞書作成手段
１０８６：モダリティ表現候補抽出手段
１０８７：モダリティ表現辞書作成手段。

【図１】

図1



【図２】

図2

コールセンター応答履歴データベースのデータ構造

1041 問合せID	1042 応答履歴メモ	1043 検索フラグ	1044 分類フラグ
1			
2			
3			
4			
5			
6			

【図３】

図3

関連シソーラス格納部のデータ構造

#	10511 タームX	10512 タームY	10513 関連度
1			
2			
3			
4			
5			
6			

【図４】

図4

タームベクトル格納部のデータ構造

#	10521 問合せID	10522 重要タームリスト
1	1	
2	2	
3	3	
4	4	
5	5	
6	6	

【図５】

図5

シソーラス概観格納部のデータ構造

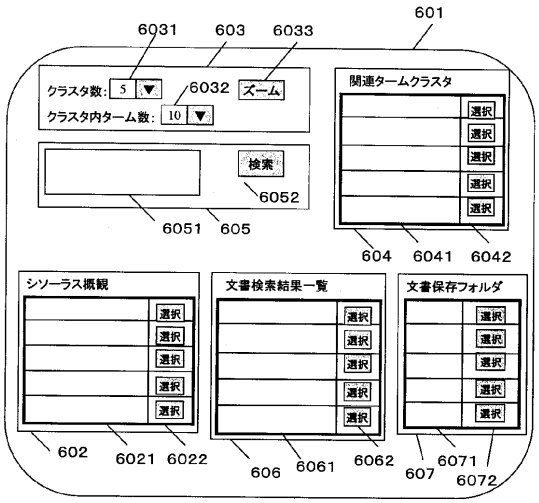
10531 10532

タームグループ番号	タームリスト
1	
2	
3	
4	
5	
6	

【図６】

図6

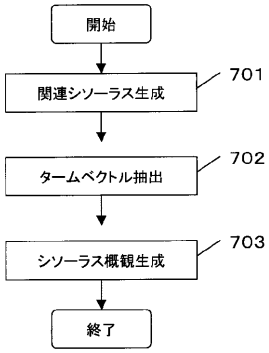
文書分類操作画面



【図７】

図7

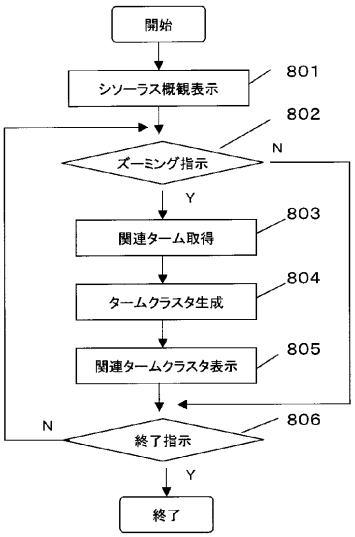
シソーラスブラウジング用データ生成処理手順



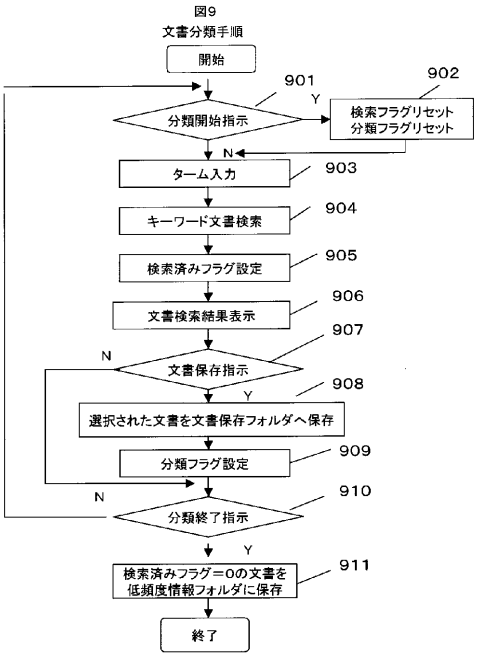
【図８】

図8

シソーラスブラウジング処理手順



【図９】



【図１０】

図10

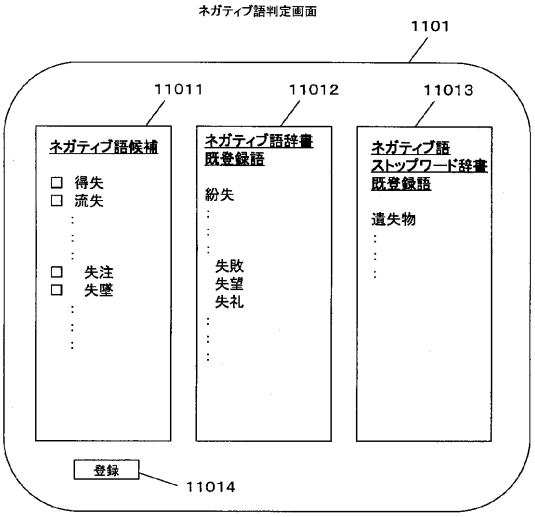
文書保存フォルダのデータ構造

1141 フォルダID	1142 フォルダ名	1143 問合せIDリスト
1		
2		
3		
4		
5		
6		

【図１１】

図11

ネガティブ語判定画面



【図１２】

図12

ネガティブ文字辞書のデータ構造

ID 10711	ネガティブ文字 10712	ネガティブ度 10713	ネガティブ語辞書登録語数 10714	ネガティブ語ストップワード辞書登録語数 10715
:				
35	失	0.8	32	4
:				
42	遅	0.7	5	0
:				
56	負	0.6	14	4

【図１３】

図13

ネガティブ語辞書のデータ構造

10721 ID	10722 ネガティブ語	10723 ネガティブ度
:		
131	紛失	0.8
132	失敗	0.8
133	失望	0.8
134	失礼	0.8
:		

【図１４】

図14

ネガティブ語ストップワード辞書のデータ構造

10731 ID	10732 ネガティブ語 ストップワード
:	
12	遺失物
:	
45	勝負
:	
:	

【図１５】

図15

モダリティ表現辞書のデータ構造

10741 ID	10742 モダリティ表現	10743 品詞	10744 モダリティ
25	くれないのか	助動詞	要望(非難)
:			
35	そもそも	副詞	批判
:			
42	なんなのか	助動詞句	質問(非難)
:			
82	欲しい	助動詞	要望

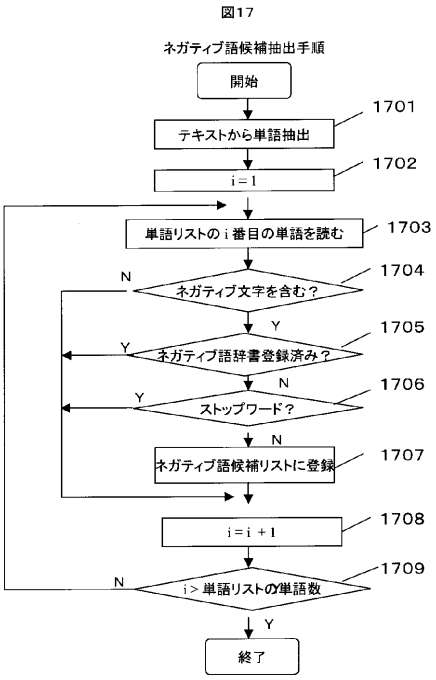
【図１６】

図16

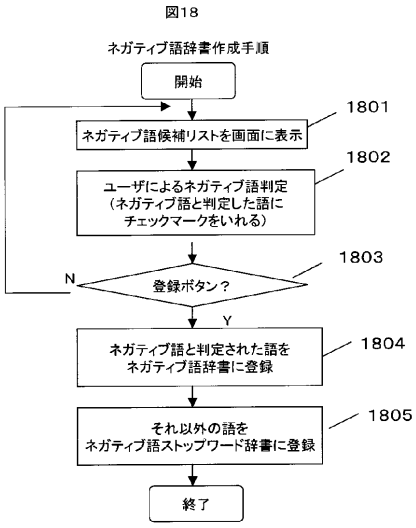
モダリティ表現ストップワード辞書のデータ構造

10751 ID	10752 モダリティ表現 ネガティブ語 ストップワード	10753 品詞
:		
12	いつも	副詞
:		
45		
:		
:		

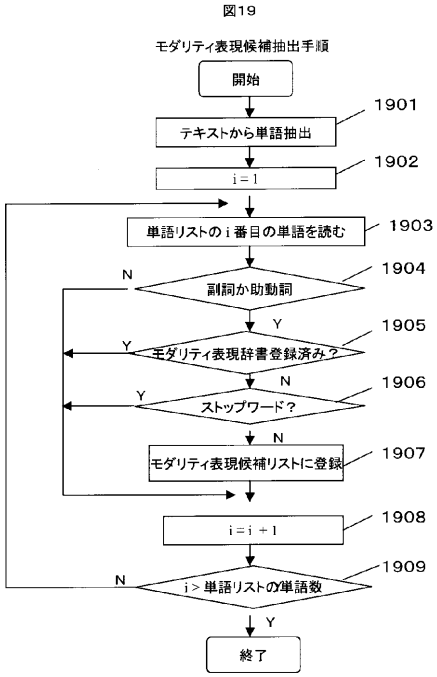
【図１７】



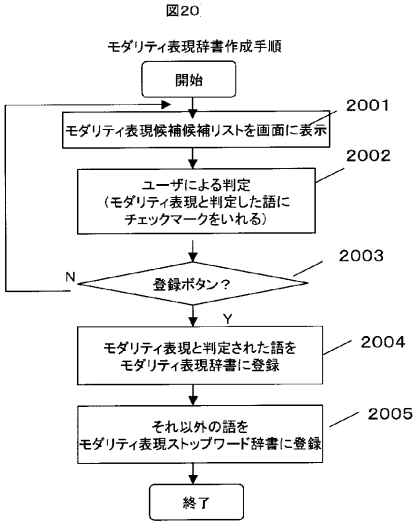
【図１８】



【図１９】



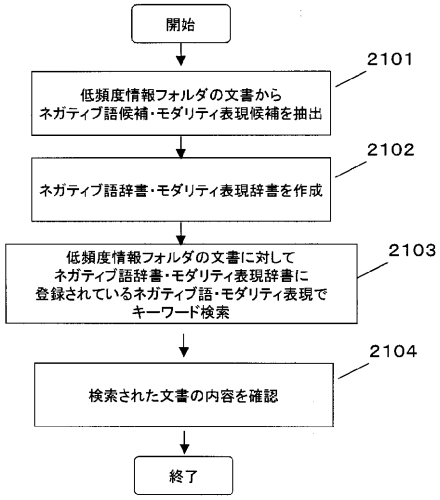
【図２０】



【図２１】

図21

低頻度知識抽出手順



フロントページの続き

(72)発明者　秋良　直人
東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所中央研究所内

Ｆターム(参考) 5B075 ND03　NK32　NR03　NR12　UU06