明　細　書

発明の名称 ：

テキストマイニングシステム、テキストマイニング方法及びプログラム

技術分野

[0001] 本発明は、テキスト情報から、その内容の傾向の分析や有用な知見を得るためのテキストマイニングシステムに関する。

背景技術

[0002] テキストマイニングとは、アンケートの自由記述欄などの自然言語で書かれたテキストデータを入力に、その内容の全体的な傾向を把握することや、有用な知見を発見するためのデータ分析手法である。例えば、コールセンターの応対メモから問い合わせの内容の把握することや商品に関するアンケートからその問題点や改善点を発見するといったことが可能となる。

[0003] 例えば、特許文献１は、２つ以上の単語の係り受け関係を抽出し、係り受けの頻度を集計し、あらかじめ与えられたカテゴリ毎に単語を整理した上で、単語間の係り受け関係のネットワークを表示している（特許文献１の図８、図１０、図１２）。例えば、カテゴリとしては、単語の品詞や主語、目的語といった文中の役割などが挙げられる。また、前記ネットワーク表示の際に、カテゴリ、単語、及び係り受け関係に関する表示条件を調節することで、広範囲な係り受け関係の全体概観から詳細な絞り込みまで自由な分析を可能にしている。

先行技術文献

特許文献

[0004] 特許文献1：特開２００７－２９３６８５号公報

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、係り受けに基づく単語間のネットワークからは、テキストの内容を直接理解しながら分析が行えないという問題がある。その理由は、

係り受け関係からはテキストの内容を正確には読み取ることができないためである。例えば、「アプリケーションをいっぱい起動させるとパソコンＢは遅い」（文１）と「パソコンＢは起動が遅い」（文２）がある場合、「パソコンＢ－遅い」と「起動－遅い」の係り受け関係が共通に含まれている。しかしながら、文１で「パソコンＢが遅い」というのは「アプリケーションをいっぱい起動させると」という条件付きであることや、文１と文２では「起動」の対象が異なっていることなど、文を関連付ける元となった係り受け関係から読み取れる情報は部分的である。そのため、テキストの具体的な内容を理解するためにはその係り受けを含む文を人間が読み、再整理する必要がある。

[0006] 本発明は、上記問題点に鑑みてなされたもので、テキストの内容を直接理解しながら分析できるテキストマイニングシステム、テキストマイニング方法、及びプログラムを提供することである。

課題を解決するための手段

[0007] 本発明は、分析対象となるテキストデータ中の各テキストから同義のテキストの集合である同義クラスタを取得する同義クラスタ取得部と、前記同義クラスタ間の含意関係を取得する含意関係取得部と、前記同義クラスタ間の含意関係から、同義クラスタを頂点とし、含意される同義クラスタから含意する同義クラスタの方向を示す有向辺を含む含意グラフを生成する含意グラフ生成部と、を備えることを特徴とするテキストマイニングシステムである。

[0008] 本発明は、分析対象となるテキストデータ中の各テキストから同義のテキストの集合である同義クラスタを取得し、前記同義クラスタ間の含意関係を取得し、前記同義クラスタ間の含意関係から、同義クラスタを頂点とし、含意される同義クラスタから含意する同義クラスタの方向を示す有向辺を含む含意グラフを生成することを特徴とする備えるテキストマイニング方法である。

[0009] 本発明は、コンピュータに、分析対象となるテキストデータ中の各テキス

トから同義のテキストの集合である同義クラスタを取得する同義クラスタ取得処理、前記同義クラスタ間の含意関係を取得する含意関係取得処理、前記同義クラスタ間の含意関係から、同義クラスタを頂点とし、含意される同義クラスタから含意する同義クラスタの方向を示す有向辺を含む含意グラフを生成する含意グラフ生成処理、を実行させることを特徴とするプログラムである。

[0010] 本発明では、テキストをそのままの形で扱い、同義クラスタ取得部が、表現が異なる同じ内容のテキストから構成されるクラスタ（同義クラスタと呼ぶ）を取得する。テキスト含意関係取得部が、前記生成した全同義クラスタ間の含意関係を取得する。ここで、含意関係として、第一のテキストの内容から第二のテキストの内容が読み取れる場合に、第一のテキストが第二のテキストを含意すると定める。本発明の構成によれば、既に同義クラスタ取得部によって同義のテキストは集約されているため、含意する側は含意される側の内容と差分の内容を含むことになる。含意グラフ生成部は、同義クラスタを頂点として、含意される側から含意する側に有向辺を引いた含意グラフを生成する。含意グラフ中の頂点は、元のテキストそのままの状態となっている。また、この含意グラフの有向きの方向、すなわち含意される側（始点）から含意する側（終点）に辿ることで、その始点のテキストの内容を含むテキストがどれくらいあり、その具体的なテキストにはどのようなものがあるかが分析できる。

発明の効果

[0011] 本発明によれば、テキストの内容を直接理解しながら分析できる。

図面の簡単な説明

[0012] [図1]図１は本発明の第１の実施形態の構成を示すブロック図である。

[図2]図２は本発明の第１の実施形態の動作を示す流れ図である。

[図3]図３はテキストデータ記憶部に記憶されているデータの一例を示した図である。

[図4]図４は同義クラスタ取得部の出力結果の一例を示した図である。

[図5]図５は含意グラフ生成部の出力結果の一例を示した図である。

[図6]図６は含意グラフ生成部の出力結果の一例を示した図である。

[図7]図７は含意グラフ生成部の結果を表示装置で表示した例を示した図である。

[図8]図８は本発明の第２の実施形態の構成を示すブロック図である。

[図9]図９は本発明の第２の実施形態の動作を示す流れ図である。

[図10]図１０は頻度集計部の結果を表示装置で表示した例を示した図である。

発明を実施するための形態

[0013] 以下、本発明の実施形態について図面を参照して説明する。

[0014] （第１の実施形態）

本発明の第１の実施形態について図面を参照して詳細に説明する。図１は、本発明の第１の実施形態の構成を示すブロック図である。図１を参照すると、本発明の実施の形態は、テキストデータ記憶部１と、テキストマイニング装置２と、表示装置３を備える。

[0015] テキストデータ記憶部１には、分析対象となるテキストデータが記憶されている。分析対象のテキストとは、人間がその内容が理解できる単位であり、具体的には１つ以上の文からなるテキストが望ましい。

[0016] テキストマイニング装置２は、同義クラスタ取得部２０とテキスト含意関係取得部２１と含意グラフ生成部２２を備える。

[0017] 同義クラスタ取得部２０は、テキストデータ記憶部１から分析対象となる各テキストデータを取り出し、同じ内容のテキストからなる同義クラスタを取得する。同義クラスタ取得部２０には、テキスト間の同義性を判定する任意の方法を利用することができる。

[0018] テキスト含意関係取得部２１は、同義クラスタ取得部２０が生成した同義クラスタ間の含意関係を取得する。テキスト含意関係取得部２１には任意の既存の含意判定方法を利用することができる。

[0019] 含意グラフ生成部２２は、テキスト含意関係取得部２１が行った含意判定

結果から、同義クラスタが頂点、同義クラスタ間の含意判定の方向を反映した有向辺から構成される含意グラフを生成する。

[0020] 表示装置３は、テキストマイニング装置２が生成した含意グラフを表示する。

[0021] 次に、図１及び図２のフローチャートを参照して本実施の形態の動作を具体的に説明する。

[0022] まず、同義クラスタ取得部２０が、テキストデータ記憶部１からテキストデータを読み出し、読みだしテキストデータについて、同義のテキストから構成されるクラスタを取得する処理を行う（ステップＳ１）。図３にテキストデータ記憶部１に記憶されているテキストデータの例を示す。図３の通り、テキストデータはテキストとその識別子を表すＩＤから構成されている。「Ｘ」は特定の製品名を表す。ステップＳ１には、同義のテキストをクラスタリングするための任意の既存の方法を用いることができる。本説明では、同義クラスタ取得部２０は、内容語から構成されたベクトルのコサイン類似度が０．８以上である場合に同義と見なし、また、同義同士も同義であるとしてクラスタリングを行う。例えば、ＩＤが１のテキストとＩＤが２のテキストはどちらも「（Ｘ，良い）」というベクトルとなり、類似度が１となるため同義と見なされる。また、ＩＤ２のテキストと類似度が０．８以上の他のテキストがあった場合には、仮にＩＤ１のテキストとの類似度が０．８未満であったとしても、そのテキストはＩＤ１のテキストとも同義とみなす。

[0023] 図４は同義クラスタリングによって得られた同義クラスタの一例である。同義クラスタは同義クラスタの識別子であるＩＤとその同義クラスタに属するテキストＩＤのリストと、同義クラスタＩＤのラベルから構成される。ラベルは同義クラスタ中の任意のテキストを取り出せばよい。

[0024] 次に、テキスト含意関係取得部２１は、同義クラスタ取得部２０によって得られた同義クラスタ間の含意関係を取得する処理を行う（ステップＳ２）。テキスト含意関係取得部２１には既存の任意のテキスト含意判定方法を用いることができる。本説明では、テキスト含意関係取得部２１は、第二のテ

キストの内容語のうち、第一のテキストに出現している内容語が８０％以上の被覆している場合に、第一のテキストが第二のテキストを含意したと判定する。また、テキストとしては、同義クラスタのラベルを用いる。ステップＳ２では、全同義クラスタから２つ組みを列挙し、片方を第一のテキスト、もう一方を第二のテキストとして含意関係を判定する。

[0025] 例えば、第一のテキストを同義クラスタのＩＤが１、第二のテキストを同義クラスタのＩＤが２の場合は、第一のテキストの内容語が「（Ｘ，良い）」で、第二のテキストの内容語が「（Ｘ，デザイン，良い）」となるため、内容語の被覆率は６６％（＝２／３）となり含意関係にはならない。一方、第一のテキストと第二のテキストを逆にすると、第一のテキストの内容語が「（Ｘ，デザイン，良い）」で、第二のテキストの内容語が「（Ｘ， 良い）」となるため、被覆率１００％（＝２／２）で含意関係と判定される。この処理を全ての同義クラスタの２つ組みについて行う。ここでは、各同義クラスタのラベル同士の含意判定結果を同義クラスタ間の含意判定結果としたが、同義クラスタには複数のテキストが存在するため、２つの同義クラスタに属する各テキストの組み合わせの含意判定結果を統合して同義クラスタ間の含意判定を行っても良い。そうすることで、同義クラスタのラベルのテキスト間の含意判定結果が誤ってしまう場合にでも正しい含意判定を行える可能性が上がる。

[0026] 最後に、含意グラフ生成部２２が、テキスト含意関係取得部２１によって得られた同義クラスタ間の含意判定結果から、同義クラスタを頂点、含意される同義クラスタから含意する同義クラスタへの有向辺からなる含意グラフを作成する処理を行う（ステップＳ３）。

[0027] 図５に、作成される含意グラフの一例を示す。図５の例では、含意される側から含意する側に有向辺を引いている。図５の含意グラフは、「Ｘが良い」を含意するテキストの集合と「Ｘはダメ」を含意するテキストの集合からなる２つの含意グラフがある。つまり「Ｘが良い」に含意されるテキストと「Ｘはダメ」に含意されるテキストは存在していないことを表す。例えば、

この含意グラフから、大きく「Ｘが良い」と「Ｘはダメ」という意見があることが分かり、「Ｘが良い」の意味を含む（含意する）の他の意見は、矢印を辿ることで「Ｘの液晶が良い」、「Ｘは軽くて良い」、「Ｘのデザインが良い」などがあることが分かる。

[0028] ここで、含意関係は推移律が成り立つため、含意グラフが冗長になることがある。例えば、「Ｘが良い」から「Ｘの液晶が良い」を経由して「Ｘの液晶がきれい」に辿ることもできるが、「Ｘが良い」から直接「Ｘの液晶がきれい」に辿ることもできる。この場合、テキストを概観するためには、意味の階層構造によって整理することが望ましいため、「Ｘが良い」から直接「Ｘの液晶がきれい」に辿るよりも、間に「Ｘの液晶が良い」というテキストもあることが把握できるように含意グラフを構成することが望ましい。そのため、含意グラフ生成部２２は、任意の頂点間に複数のパスが有る場合に、その最長パスのみを残して含意グラフを構成するようにしてもよい。パスの長さは、例えばそのパスに含まれる辺の数を用いることができる。その場合は、図５の左の含意グラフは図６のように生成される。例えば、「Ｘが良い」から「Ｘの液晶がきれい」に辿るパスは、直接辿れる辺と、「Ｘの液晶が良い」を経由して２つの辺を辿れるパスがある。この場合、最長パスを採用すると、直接辿れる辺が除去される。全ての頂点間でこの処理を行うと図６のように、できるだけ深い階層を持つように含意グラフを構成できる。

[0029] また、含意グラフ生成部２２は、各同義クラスタ間の含意関係について、含意らしさのスコアを算出し付与しても良い。そのスコアを用いることで、含意の可能性が高い（スコアが所定値よりも高い）辺のみを残して分析するといった処理が可能となる。この場合、パスの長さは、パスに含まれる辺のスコアの和とすることが考えられる。

[0030] 表示装置３は、このように含意グラフ生成部２２により作成された含意グラフを表示する。その表示方法としては、含意グラフ全体が膨大になるため、どの頂点からも辿れない同義クラスタを起点として表示し、その同義クラスタをクリックすることで、近傍の同義クラスタを展開表示する方法が望ま

しい。図７に図６の含意グラフの表示方法の例を示す。図６には、どの同義クラスタからも辿れない頂点が「Ｘが良い」と「Ｘがダメ」の２つがある。この時、「Ｘが良い」がクリックされると、その「Ｘが良い」から一つの辺で辿れる「Ｘの液晶が良い」、「Ｘは軽くて良い」、「Ｘのデザインが良い」を展開する表示処理を行う。ここで、一つの辺ではなく、所定個数の辺で辿れる近傍を展開する方法や、一定数以下の同義クラスタを表示するという制限を付け、その制限の範囲で近傍を展開するという表示処理を行っても良い。このような表示制御処理は、テキストマイニング装置２の含意グラフ生成部２２が行ってもよく、また、表示装置３が行ってもよい。

[0031] 以上の第１の実施の形態の発明によって、テキストの内容を直接理解しながら分析ができるようになる。その理由は、テキストをそのままの形で扱い、頂点を同義クラスタ、有向辺を含意関係とする含意グラフを生成していることから、頂点がテキストの情報をそのまま表し、また、その内容を実質的に含む他のテキストを含意グラフによって辿ることができるためである。

[0032] （第２の実施形態）

次に、本発明の第２の実施の形態について図８を用いて説明する。図８を参照すると、本発明の第２の実施の形態は、第１の実施の形態の発明のテキストマイニング装置に、さらに頻度集計部４０を備える。

[0033] 頻度集計部４０は、第１の実施の形態の発明によって生成された含意グラフを用いて、各同義クラスタが表す情報の頻度を、各同義クラスタから含意グラフによって辿れる同義クラスタの情報の頻度も用いて集計する処理を行う。

[0034] 図８及び図９のフローチャートを参照して本実施の形態の動作を具体的に説明する。

[0035] 図８のステップＳ１～Ｓ３は第１の実施の形態である図２のフローチャートと同様である。頻度集計部４０は、含意グラフの各頂点の同義クラスタの情報の頻度を、その頂点の同義クラスタと、その頂点から移動できる頂点の同義クラスタとに含まれるテキストのテキストＩＤの数として集計する（ス

テップＳ４）。これは、例えば、各頂点から、含意グラフを深さ優先で探索して推移可能な頂点集合を列挙し、各頂点の同義クラスタのテキストＩＤの数の和を取ることで可能となる。他にも、ある頂点から推移可能な頂点集合を列挙できる任意のアルゴリズムを用いることができる。

[0036] 表示装置３は、頻度集計部４０による頻度集計結果を表示する。基本的には、図７の各同義クラスタノード上に頻度集計結果を表示する方法が考えられる。また、表形式で表示する方法も考えられる。図１０に図７の含意グラフによる頻度集計結果を表形式で表示した例を示す。まず、「Ｘは良い」、「Ｘはダメ」の２つの同義クラスタがあり、それぞれの同義クラスタの頻度が表示されている。「Ｘは良い」がクリックされると、そこから一つの辺で辿れる同義クラスタとその頻度を表として表示する処理を行う。このような表示制御処理は、テキストマイニング装置２の含意グラフ生成部２２や頻度集計部４０が行ってもよく、また、表示装置３が行ってもよい。

[0037] 以上の第２の実施の形態の発明によって、各テキストの意味が実質的に含まれるテキストも含めて頻度を集計することが可能となる。その理由は、第１の実施の形態の発明によって、含意関係による含意グラフを生成し、その含意グラフによって各頂点のテキストの情報を実質的に含む、すなわち含意する同義クラスタを列挙できるためである。

[0038] 上述した本発明の実施形態に係るテキストマイニング装置は、ＣＰＵ（Central Processing Unit）が記憶部に格納された動作プログラム等を読み出して実行することにより実現されてもよく、また、ハードウェアで構成されてもよい。上述した実施の形態の一部の機能のみをコンピュータプログラムにより実現することもできる。

[0039] 上記の実施形態の一部又は全部は、以下の付記のようにも記載されうるが、以下には限られない。

[0040] （付記１）

分析対象となるテキストデータ中の各テキストから同義のテキストの集合である同義クラスタを取得する同義クラスタ取得部と、

前記同義クラスタ間の含意関係を取得する含意関係取得部と、

前記同義クラスタ間の含意関係から、同義クラスタを頂点とし、含意される同義クラスタから含意する同義クラスタの方向を示す有向辺を含む含意グラフを生成する含意グラフ生成部と、

を備えることを特徴とするテキストマイニングシステム。

[0041] （付記２）

前記含意グラフ生成部が、２つの前記同義クラスタ間に複数のパスがある場合、最長パスを構成する辺のみを残して含意グラフを作成する

ことを特徴とする付記１に記載のテキストマイニングシステム。

[0042] （付記３）

前記各同義クラスタが表す情報の頻度を、前記含意グラフにおいてその当該同義クラスタから辿れる同義クラスタの情報の頻度も用いて集計する頻度集計部をさらに備える

ことを特徴とする付記１又は付記２に記載のテキストマイニングシステム。

[0043] （付記４）

前記含意グラフを用いて、各同義クラスタの分析時に、当該含意グラフ上の近傍の同義クラスタを表示する表示装置をさらに備える

ことを特徴とする付記１から付記３のいずれかに記載のテキストマイニングシステム。

[0044] （付記５）

分析対象となるテキストデータ中の各テキストから同義のテキストの集合である同義クラスタを取得し、

前記同義クラスタ間の含意関係を取得し、

前記同義クラスタ間の含意関係から、同義クラスタを頂点とし、含意される同義クラスタから含意する同義クラスタの方向を示す有向辺を含む含意グラフを生成する

ことを特徴とする備えるテキストマイニング方法。

[0045] （付記６）

２つの前記同義クラスタ間に複数のパスがある場合、最長パスを構成する辺のみを残して含意グラフを作成する

ことを特徴とする付記５に記載のテキストマイニング方法。

[0046] （付記７）

前記各同義クラスタが表す情報の頻度を、前記含意グラフにおいてその当該同義クラスタから辿れる同義クラスタの情報の頻度も用いて集計する

ことを特徴とする付記５又は付記６に記載のテキストマイニング方法。

[0047] （付記８）

前記含意グラフを用いて、各同義クラスタの分析時に、当該含意グラフ上の近傍の同義クラスタを表示する

ことを特徴とする付記５から付記７のいずれかに記載のテキストマイニング方法。

[0048] （付記９）

コンピュータに、

分析対象となるテキストデータ中の各テキストから同義のテキストの集合である同義クラスタを取得する同義クラスタ取得処理、

前記同義クラスタ間の含意関係を取得する含意関係取得処理、

前記同義クラスタ間の含意関係から、同義クラスタを頂点とし、含意される同義クラスタから含意する同義クラスタの方向を示す有向辺を含む含意グラフを生成する含意グラフ生成処理、

を実行させることを特徴とするプログラム。

[0049] （付記１０）

前記含意グラフ生成処理は、２つの前記同義クラスタ間に複数のパスがある場合、最長パスを構成する辺のみを残して含意グラフを作成する

ことを特徴とする付記９に記載のプログラム。

[0050] （付記１１）

前記コンピュータに、

前記各同義クラスタが表す情報の頻度を、前記含意グラフにおいてその当

該同義クラスタから辿れる同義クラスタの情報の頻度も用いて集計する頻度集計処理をさらに実行させる

ことを特徴とする付記９又は付記１０に記載のプログラム。

[0051] （付記１２）

前記コンピュータに、

前記含意グラフを用いて、各同義クラスタの分析時に、当該含意グラフ上の近傍の同義クラスタを表示する処理を実行させる

ことを特徴とする付記９から付記１１のいずれかに記載のプログラム。

[0052] 以上、好ましい実施の形態をあげて本発明を説明したが、本発明は必ずしも上記実施の形態に限定されるものではなく、その技術的思想の範囲内において様々に変形し実施することが出来る。

この出願は、２０１２年４月２６日に出願された日本出願特願２０１２－１０１０１７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

[0053]

１	テキストデータ記憶部
２	テキストマイニング装置
２０	同義クラスタ取得部
２１	テキスト含意関係取得部
２２	含意グラフ生成部
３	表示装置
４	テキストマイニング装置
４０	頻度集計部

請求の範囲

［請求項1］ 分析対象となるテキストデータ中の各テキストから同義のテキストの集合である同義クラスタを取得する同義クラスタ取得部と、

前記同義クラスタ間の含意関係を取得する含意関係取得部と、

前記同義クラスタ間の含意関係から、同義クラスタを頂点とし、含意される同義クラスタから含意する同義クラスタの方向を示す有向辺を含む含意グラフを生成する含意グラフ生成部と、

を備えることを特徴とするテキストマイニングシステム。

［請求項2］ 前記含意グラフ生成部が、２つの前記同義クラスタ間に複数のパスがある場合、最長パスを構成する辺のみを残して含意グラフを作成する

ことを特徴とする請求項１に記載のテキストマイニングシステム。

［請求項3］ 前記各同義クラスタが表す情報の頻度を、前記含意グラフにおいてその当該同義クラスタから辿れる同義クラスタの情報の頻度も用いて集計する頻度集計部をさらに備える

ことを特徴とする請求項１又は請求項２に記載のテキストマイニングシステム。

［請求項4］ 前記含意グラフを用いて、各同義クラスタの分析時に、当該含意グラフ上の近傍の同義クラスタを表示する表示装置をさらに備える

ことを特徴とする請求項１から請求項３のいずれかに記載のテキストマイニングシステム。

［請求項5］ 分析対象となるテキストデータ中の各テキストから同義のテキストの集合である同義クラスタを取得し、

前記同義クラスタ間の含意関係を取得し、

前記同義クラスタ間の含意関係から、同義クラスタを頂点とし、含意される同義クラスタから含意する同義クラスタの方向を示す有向辺を含む含意グラフを生成する

ことを特徴とする備えるテキストマイニング方法。

[請求項6] コンピュータに、

分析対象となるテキストデータ中の各テキストから同義のテキストの集合である同義クラスタを取得する同義クラスタ取得処理、

前記同義クラスタ間の含意関係を取得する含意関係取得処理、

前記同義クラスタ間の含意関係から、同義クラスタを頂点とし、含意される同義クラスタから含意する同義クラスタの方向を示す有向辺を含む含意グラフを生成する含意グラフ生成処理、

を実行させることを特徴とするプログラム。