

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第6085888号

(P6085888)

(45) 発行日 平成29年3月1日(2017.3.1)

(24) 登録日 平成29年2月10日(2017.2.10)

(51) Int. Cl.

F 1

G06F 17/30 (2006.01)

G06F 17/30 220Z

G06F 17/27 (2006.01)

G06F 17/30 170A

G06Q 30/02 (2012.01)

G06F 17/27 615

G06Q 30/02 312

請求項の数 7 (全 21 頁)

(21) 出願番号 特願2014-174500 (P2014-174500)
 (22) 出願日 平成26年8月28日(2014.8.28)
 (65) 公開番号 特開2016-51220 (P2016-51220A)
 (43) 公開日 平成28年4月11日(2016.4.11)
 審査請求日 平成28年4月28日(2016.4.28)

(73) 特許権者 301017433
 有限責任監査法人トーマツ
 東京都港区港南二丁目15番3号 品川イ
 ンターシティ
 (74) 代理人 100101236
 弁理士 栗原 浩之
 (74) 代理人 100166914
 弁理士 山▲崎▼ 雄一郎
 (72) 発明者 野守 耕爾
 東京都千代田区丸の内3-3-1 有限責
 任監査法人トーマツ内
 (72) 発明者 神津 友武
 東京都千代田区丸の内3-3-1 有限責
 任監査法人トーマツ内

最終頁に続く

(54) 【発明の名称】 分析方法、分析装置及び分析プログラム

(57) 【特許請求の範囲】

【請求項1】

テキストデータ、及び当該テキストデータに関する属
 性情報を分析する分析装置が、

前記テキストデータから文章を抽出し、各文章から、
 予め定めた第1品詞及び第2品詞のそれぞれに該当する
 第1単語群及び第2単語群を抽出し、各文章に含まれて
 いる第1単語群に属する単語及び第2単語群に属する単
 語の組み合わせの個数を表す共起行列を作成する単語抽
 出ステップと、

前記共起行列を入力とし、第1単語群に属する単語及
 び第2単語群に属する単語で構成される複数のクラスを
 抽出する潜在意味解析法を実行することにより、各クラ
 スを条件とした第1単語群に属する単語の第1条件付確
 率、及び各クラスを条件とした第2単語群に属する単語

の第2条件付確率を求めるクラス抽出ステップと、

前記第1条件付確率及び第1単語群の発生数、並びに
 前記第2条件付確率及び第2単語群の発生数に基づいて
 、各クラスを条件とした各文章の条件付確率を計算し、
 各テキストデータに対する各クラスのスコアを求めるス
 コア計算ステップと、

前記クラス及び前記属性情報を変数として、モデル化
 手法を用いてそれらの関係をモデル化するモデル化ステ
 ップと、

10 前記クラスの変数を変化させたときの前記属性情報の
 状態、又は前記属性情報の変数を変化させた時の前記ク
 ラスの状態を推論する推論ステップと、
 を実行することを特徴とする分析方法。

【請求項2】

請求項1に記載する分析方法において、

前記潜在意味解析法は、確率的潜在意味解析（PLS A: Probabilistic Latent Semantic Analysis）法である

ことを特徴とする分析方法。

【請求項3】

請求項2に記載する分析方法において、

前記分析装置は、前記確率的潜在意味解析においては、抽出するクラスの数を変えて複数回実行し、各計算結果における情報量基準を算出し、当該情報量基準が最適となる計算結果のときの前記第1条件付確率及び第2条件付確率を求める

ことを特徴とする分析方法。

【請求項4】

請求項1～請求項3の何れか一項に記載する分析方法において、

前記分析装置は、前記スコア計算ステップでは、前記第1条件付確率及び第1単語群の発生数、並びに前記第2条件付確率及び第2単語群の発生数に基づいて各クラスを条件とした各文章の条件付確率を計算し、当該条件付確率を各文章の発生確率で除した値を、各文章に対する各クラスのスコアとし、当該スコアをテキストデータ単位に集約することで各テキストデータに対する各クラスのスコアを決定する

ことを特徴とする分析方法。

【請求項5】

請求項1～請求項4の何れか一項に記載する分析方法において、

前記モデル化手法は、ベイジアンネットワークであることを特徴とする分析方法。

【請求項6】

テキストデータ、及び当該テキストデータに関する属性情報の分析装置であって、

前記テキストデータから文章を抽出し、各文章から、予め定めた第1品詞及び第2品詞のそれぞれに該当する第1単語群及び第2単語群を抽出し、各文章に含まれている第1単語群に属する単語及び第2単語群に属する単語の組み合わせの個数を表す共起行列を作成する単語抽出手段と、

前記共起行列を入力とし、第1単語群に属する単語及び第2単語群に属する単語で構成される複数のクラスを抽出する潜在意味解析法を実行することにより、各クラスを条件とした第1単語群に属する単語の第1条件付確率、及び各クラスを条件とした第2単語群に属する単語の第2条件付確率を求めるクラス抽出手段と、

前記第1条件付確率及び第1単語群の発生数、並びに前記第2条件付確率及び第2単語群の発生数に基づいて、各クラスを条件とした各文章の条件付確率を計算し、各テキストデータに対する各クラスのスコアを求めるスコア計算手段と、

前記クラス及び前記属性情報を変数として、モデル化

手法を用いてそれらの関係をモデル化するモデル化手段と、

前記クラスの変数を変化させたときの前記属性情報の状態、又は前記属性情報の変数を変化させた時の前記クラスの状態を推論する推論手段を備える

ことを特徴とする分析装置。

【請求項7】

テキストデータ、及び当該テキストデータに関する属性情報をコンピュータに分析させる分析プログラムであって、

前記コンピュータを、

前記テキストデータから文章を抽出し、各文章から、予め定めた第1品詞及び第2品詞のそれぞれに該当する第1単語群及び第2単語群を抽出し、各文章に含まれている第1単語群に属する単語及び第2単語群に属する単語の組み合わせの個数を表す共起行列を作成する単語抽出手段と、

前記共起行列を入力とし、第1単語群に属する単語及び第2単語群に属する単語で構成される複数のクラスを抽出する潜在意味解析法を実行することにより、各クラスを条件とした第1単語群に属する単語の第1条件付確率、及び各クラスを条件とした第2単語群に属する単語の第2条件付確率を求めるクラス抽出手段と、

前記第1条件付確率及び第1単語群の発生数、並びに前記第2条件付確率及び第2単語群の発生数に基づいて、各クラスを条件とした各文章の条件付確率を計算し、各テキストデータに対する各クラスのスコアを求めるスコア計算手段と、

前記クラス及び前記属性情報を変数として、モデル化手法を用いてそれらの関係をモデル化するモデル化手段と、

前記クラスの変数を変化させたときの前記属性情報の状態、又は前記属性情報の変数を変化させた時の前記クラスの状態を推論する推論手段として機能させるための分析プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、テキストデータの分析を行う場合に、データに記載されている内容の現状を把握するだけでなく、条件を変化させたときにどのような結果となりうるのか推論する分析方法、分析装置及び分析プログラムに関する。

【背景技術】

【0002】

テキストの電子化の急増とテキストマイニングツールの普及に伴い、テキストデータからいかに有用な知識を抽出するかということが課題となっている。テキストマイニングは非構造化データであるテキストを統計的に分析可能な形にする自然言語処理技術であり、これによ

て大量の文章からどんなことが発言・記録されているのか定量的に分析を進めることができる（非特許文献1参照）。近年ではその適用事例も増えてきており、コールセンターの対応履歴や顧客満足度調査の自由記述回答、営業日報、Web上の書き込みなど、様々な分野で適用され経営に活用されている（非特許文献2～4参照）。

【0003】

テキストデータの分析でよく実行されることは、テキストマイニングを適用して統計処理可能になったデータを用いて、出現単語の頻度集計をして全体像を把握したり、他の属性情報別の集計をしてその対応関係を分析したり、時系列変化の特徴を把握したり、あるいは文章をクラスタリングすることなどがあり、現状を把握する手段として有用である。

【0004】

一方、テキストデータとその属性データを用いて、そこで記述されている状況を様々な条件の下で推論可能にする分析方法が提案されている。例えば、テキストデータからマイニングされた単語情報を変数とし、ベイジアンネットワークによりモデル化する事例がある（非特許文献5参照）。ベイジアンネットワークは複数の変数の間の依存関係をグラフ構造によって表わし、その変数間の定量的な関係を条件付き確率によって表わした確率的なモデリング手法であり、ある変数の値が観測されたときに、未観測の変数の確率分布を推論することができる（非特許文献6参照）。

【0005】

非特許文献5では、抽出された単語一つ一つをそのまま変数としているためモデルが非常に複雑で使いにくいものとなっており、またベイジアンネットワークのモデルのベースとなる条件付確率表も疎になりやすく、正しい推論ができない可能性が生じてしまう。

【0006】

単語そのものではなく文章のトピックを抽出する手法として、PLSAがある（非特許文献7参照）。元々文章分類のために開発された手法で、文章とそこに出現する単語の間には観測できない潜在的な意味クラスがあることを想定し、文章と単語の共通のトピックとなるような特徴を見つける手法である。このような手法により抽出されたトピックを変数として扱い、ベイジアンネットワークでモデルを構築することで、モデルがシンプルとなり、結果の解釈もしやすくなる可能性がある。

【0007】

PLSAとベイジアンネットワークの応用例としては、テキストデータではないが、購買データにPLSAを適用し、PLSAによって抽出されたクラスを変数として扱い、ベイジアンネットワークを適用して他の変数間との関係をモデル化する事例がある（非特許文献8参照）。

【先行技術文献】

【非特許文献】

【0008】

【非特許文献1】 那須川哲哉:テキストマイニングを使う技術/作る技術:基礎技術と適用事例から導く本質と活用法, 東京電機大学出版局, 2006.

【非特許文献2】 長谷川久:テキストマイニングの利用による早期人材育成の実践-コール・ログ分析による要員育成の効率化-, 情報処理学会デジタルプラクティス, Vol. 2, No. 3, pp. 192-199, 2011.

【非特許文献3】 市村由美, 鈴木優, 酢山明弘, 折原良平, 中山康子:日報分析システムと分析用知識記述支援ツールの開発, 電子情報通信学会論文誌D-2, Vol. J86-D-2, No. 2, pp. 310-323, 2003.

【非特許文献4】 三川健太, 高橋勉, 後藤正幸:テキストデータに基づく顧客ロイヤルティの構造分析手法に関する一考察, 日本経営工学会論文誌, Vol. 58, No. 3, pp. 182-192, 2007.

【非特許文献5】 野守耕爾, 北村光司, 本村陽一, 西田佳史, 山中龍宏, 小松原明哲:大規模傷害テキストデータに基づいた製品に対する行動と事故の関係モデルの構築:エビデンスベースド・リスクアセスメントの実現に向けて, 人工知能学会論文誌, Vol. 25, No. 5, pp. 602-612, 2010.

【非特許文献6】 Motomura, Y.:BAYONET, Bayesian Network on Neural Network, Foundation of Real-World Intelligence, pp.28-37, 2001.

【非特許文献7】 Hofmann, T.:Probabilistic latent semantic analysis, Proc. Of Uncertainty in Artificial Intelligence, pp.289-296, 1999.

【非特許文献8】 石垣司, 竹中毅, 本村陽一:百貨店ID付きPOSデータからのカテゴリ別状況依存の変数間関係の自動抽出法, オペレーションズ・リサーチ:経営の科学, Vol. 56, No. 2, pp. 77-83, 2011.

【発明の概要】

【発明が解決しようとする課題】

【0009】

上述したように、単語や係り受け関係の頻度を集計する手法は、分析対象のテキストデータそれ自体の中身の把握をして、改善すべき点やニーズを抽出する上では分かりやすく有効な手段であるが、現状把握に留まるアプローチといえる。例えば条件を変化させたとき、改善を施したときに、その結果がどのように変化するのかシミュレーションすることはできない。

【0010】

また、テキストデータから単語を抽出して、その単語を変数としてベイジアンネットワークによるモデルを構築した場合、出現単語が変化したときの結果をシミュレーションできるが、モデルが複雑であり、また、正しい推論ができない可能性がある。

【0011】

さらに、PLSAは、文章のトピックを抽出するにあたり、文章と単語との関係からクラスを導きだすことができるが、そのクラスの意味付け（トピックの決定）は人間が行う必要がある。しかし、そのクラスに属するとされた文章と、それに関係する単語とを解釈して、適切なトピックを決定することが困難な場合もある。

【0012】

本発明は、上記事情に鑑みてなされたものであり、テキストデータに含まれる文章についてトピックの抽出を容易とし、当該トピックに基づいてベイジアンネットワークによるモデル化をすることで、モデルが複雑になることを回避し、さらに、そのモデル化結果において、条件を変化させたときにどのような結果となりうるのかを推論することができる分析方法、分析装置及び分析プログラムを提供することを目的とする。

【課題を解決するための手段】

【0013】

上記課題を解決する本発明の第1の態様は、テキストデータ、及び当該テキストデータに関する属性情報を分析する分析装置が、前記テキストデータから文章を抽出し、各文章から、予め定めた第1品詞及び第2品詞のそれぞれに該当する第1単語群及び第2単語群を抽出し、各文章に含まれている第1単語群に属する単語及び第2単語群に属する単語の組み合わせの個数を表す共起行列を作成する単語抽出ステップと、前記共起行列を入力とし、第1単語群に属する単語及び第2単語群に属する単語で構成される複数のクラスを抽出する潜在意味解析法を実行することにより、各クラスを条件とした第1単語群に属する単語の第1条件付確率、及び各クラスを条件とした第2単語群に属する単語の第2条件付確率を求めるクラス抽出ステップと、前記第1条件付確率及び第1単語群の発生数、並びに前記第2条件付確率及び第2単語群の発生数に基づいて、各クラスを条件とした各文章の条件付確率を計算し、各テキストデータに対する各クラスのスコアを求めるスコア計算ステップと、前記クラス及び前記属性情報を変数として、モデル化手法を用いてそれらの関係をモデル化するモデル化ステップと、前記クラスの変数を変化させたときの前記属性情報の状態、又は前記属性情報の変数を変化させた時の前記クラスの状態を推論する推論ステップと、を実行することを特徴とする分析方法にある。

【0014】

前記テキストデータには、複数の文章が含まれることがあり、本発明でいう文章とは、テキストデータに含まれる一文である。

【0015】

本発明の第2の態様は、第1の態様に記載する分析方法において、前記潜在意味解析法は、確率的潜在意味解析(PLSA: Probabilistic Latent Semantic Analysis)法である

ことを特徴とする分析方法にある。

【0016】

本発明の第3の態様は、第2の態様に記載する分析方法において、前記分析装置は、前記確率的潜在意味解析においては、抽出するクラスの数を変えて複数回実行し、各計算結果における情報量基準を算出し、当該情報量基準が最適となる計算結果のときの前記第1条件付確率及び第2条件付確率を求めることを特徴とする分析方法にある。

10 【0017】

本発明の第4の態様は、第1～第3の何れか一つの態様に記載する分析方法において、前記分析装置は、前記スコア計算ステップでは、前記第1条件付確率及び第1単語群の発生数、並びに前記第2条件付確率及び第2単語群の発生数に基づいて各クラスを条件とした各文章の条件付確率を計算し、当該条件付確率を各文章の発生確率で除した値を、各文章に対する各クラスのスコアとし、当該スコアをテキストデータ単位に集約することで各テキストデータに対する各クラスのスコアを決定することを特徴とする分析方法にある。

20 【0018】

本発明の第5の態様は、第1～第4の何れか一つの態様に記載する分析方法において、前記モデル化手法は、ベイジアンネットワークであることを特徴とする分析方法にある。

【0019】

本発明の第6の態様は、テキストデータ、及び当該テキストデータに関する属性情報の分析装置であって、前記テキストデータから文章を抽出し、各文章から、予め定めた第1品詞及び第2品詞のそれぞれに該当する第1単語群及び第2単語群を抽出し、各文章に含まれている第1単語群に属する単語及び第2単語群に属する単語の組み合わせの個数を表す共起行列を作成する単語抽出手段と、前記共起行列を入力とし、第1単語群に属する単語及び第2単語群に属する単語で構成される複数のクラスを抽出する潜在意味解析法を実行することにより、各クラスを条件とした第1単語群に属する単語の第1条件付確率、及び各クラスを条件とした第2単語群に属する単語の第2条件付確率を求めるクラス抽出手段と、前記第1条件付確率及び第1単語群の発生数、並びに前記第2条件付確率及び第2単語群の発生数に基づいて、各クラスを条件とした各文章の条件付確率を計算し、各テキストデータに対する各クラスのスコアを求めるスコア計算手段と、前記クラス及び前記属性情報を変数として、モデル化手法を用いてそれらの関係をモデル化するモデル化手段と、前記クラスの変数を変化させたときの前記属性情報の状態、又は前記属性情報の変数を変化させた時の前記クラスの状態を推論する推論手段を備えることを特徴とする分析装置にある。

50 【0020】

本発明の第7の態様は、テキストデータ、及び当該テキストデータに関する属性情報をコンピュータに分析させる分析プログラムであって、前記コンピュータを、前記テキストデータから文章を抽出し、各文章から、予め定めた第1品詞及び第2品詞のそれぞれに該当する第1単語群及び第2単語群を抽出し、各文章に含まれている第1単語群に属する単語及び第2単語群に属する単語の組み合わせの個数を表す共起行列を作成する単語抽出手段と、前記共起行列を入力とし、第1単語群に属する単語及び第2単語群に属する単語で構成される複数のクラスを抽出する潜在意味解析法を実行することにより、各クラスを条件とした第1単語群に属する単語の第1条件付確率、及び各クラスを条件とした第2単語群に属する単語の第2条件付確率を求めるクラス抽出手段と、前記第1条件付確率及び第1単語群の発生数、並びに前記第2条件付確率及び第2単語群の発生数に基づいて、各クラスを条件とした各文章の条件付確率を計算し、各テキストデータに対する各クラスのスコアを求めるスコア計算手段と、前記クラス及び前記属性情報を変数として、モデル化手法を用いてそれらの関係をモデル化するモデル化手段と、前記クラスの変数を変化させたときの前記属性情報の状態、又は前記属性情報の変数を変化させた時の前記クラスの状態を推論する推論手段として機能させるための分析プログラムにある。

【発明の効果】

【0021】

本発明によれば、テキストデータに含まれる文章についてトピックの抽出を容易とし、当該トピックに基づいてベイジアンネットワークによるモデル化をすることで、モデルが複雑になることを回避し、さらに、そのモデル化結果において、条件を変化させたときにどのような結果となりうるのかを推論することができる分析方法、分析装置及び分析プログラムが提供される。

【図面の簡単な説明】

【0022】

【図1】本実施形態に係る分析方法を実行する分析プログラムを実行する分析装置の機能ブロック図である。

【図2】PLSAの概念図である。

【図3】モデル化手段により得られたベイジアンネットワークの一部である。

【図4】分析装置での処理を示すフローチャートである。

【発明を実施するための形態】

【0023】

以下、本発明を実施するための形態について説明する。なお、実施形態の説明は例示であり、本発明は以下の説明に限定されない。

【0024】

〈実施形態1〉

図1は、本実施形態に係る分析方法を実行する分析プログラムを実行する分析装置の機能ブロック図である。分析プログラム10は、分析装置1にインストールされて実行されるものである。分析装置1は、特に図示しないが、CPU、RAM、ハードディスク、入出力装置、通信手段等を備えた一般的なコンピュータである。

【0025】

ハードディスクには、分析装置1のCPU等を制御するためのオペレーティングシステムがインストールされている。このオペレーティングシステムにより、ハードディスクにインストールされた分析プログラム10がRAMに読み込まれ、RAMに読み込まれた分析プログラムがCPUにより実行される。

【0026】

このような分析プログラムの処理対象となるテキストデータ及び属性情報について説明する。テキストデータとは、文章を符号化したデータである。もちろん、符号化の方式（文字コード）は特に限定はなく、符号化により表される言語の種別も問わない。本実施形態では、テキストデータは、日本語の文からなり、UTF-8など文字コードで表現されている。属性情報とは、テキストデータに関する情報である。

【0027】

本実施形態では、テキストデータとして、宿泊施設を利用した利用者が、その宿泊施設を利用したときの感想など、いわゆるクチコミを例に挙げる。属性情報としては、利用者属性、施設属性、宿泊内容、項目得点を例に挙げる。利用者属性とは、年齢、性別などの利用者に関する情報である。施設属性とは、利用者が利用した施設が備えるラウンジ、LANなどの設備の有無など施設に関する情報である。宿泊内容とは、利用者が宿泊施設を利用した際の宿泊料金や部屋サイズ、食事の有無など宿泊に関する情報である。項目得点とは、利用者が宿泊施設に付けた点数であり、総合得点、朝食について評価した得点（朝食得点）、接客対応などについて評価した得点（サービス得点）、風呂設備について評価した得点（風呂得点）などが挙げられる。

【0028】

このようなテキストデータは、例えば、インターネット上の宿泊施設に関する評価サイトや、宿泊施設の紹介、予約仲介サイトなどで、利用者から入力されたものである。利用者は、利用した宿泊施設について意見や感想などを、非定型のテキストデータとして入力するとともに、宿泊施設に関する情報や評価を定型的な属性情報として入力する。表1にテキストデータ及びテキストデータのそれぞれに関連づけられた属性情報の一例を示す。

【0029】

【表1】

テキスト データID	テキストデータ	属性情報												
		利用者属性		施設属性		宿泊内容		項目得点						
		性別	年代	ラウンジ	LAN	宿泊料金	部屋サイズ	総合得点	朝食得点	夕食得点	サービス得点	部屋得点	風呂得点	清潔得点
1	スタッフがものすごく丁寧で、部屋も広く綺麗でした。また、食事は豪華で美味しかったです。	男	20代	なし	あり	12,000	25m ²	4	3	2	5	4	2	3
2	朝食はバイキング形式でボリュームもあり満足でした。部屋が少し狭いのが残念。	女	30代	あり	あり	9,800	18m ²	3	4	2	3	2	3	3
3	部屋は景色がよかった。風呂も快適だった。	男	50代	あり	あり	15,000	29m ²	5	4	4	3	5	4	4

【0030】

テキストデータIDは、個々のテキストデータを識別する情報であり、ここでは重複しない数値である。各テキストデータには属性情報が関連づけられている。なお、利用者属性、施設属性、宿泊内容、項目得点は、表1には示したものに限定されず、任意の項目を設定可能である。

【0031】

このようなテキストデータ及び属性情報を分析対象とする分析装置1は、単語抽出手段11、クラス抽出手段12、スコア計算手段13、モデル化手段14、推論手段15とを備えている。本実施形態では、それらの各手段は、分析装置1で実行される分析プログラム10として実装されている。すなわち、分析プログラム10は、分析装置1を各手段11～15として機能させるプログラムである。

【0032】

30 単語抽出手段11は、テキストデータから文章を抽出し、各文章から、予め定めた第1品詞及び第2品詞のそれぞれに該当する第1単語群及び第2単語群を抽出し、各文章に含まれている第1単語群に属する単語及び第2単語群に属する単語の組み合わせの個数を表す共起行列を作成する。

【0033】

テキストデータには、複数の文章が含まれることがあり、本発明でいう文章とは、テキストデータに含まれる一文である。分析装置1で実行される分析プログラム10の単語抽出手段11は、テキストデータの一つずつ読み込み、これを句点や「?」「!」など一文の末尾に用いられる文字を基準として個別の文章として出力する。例えば、テキストデータID「1」については、次のように2つの文章が抽出される。

【0034】

【表2】

テキストデータID	文章ID	テキストデータ
1	1	スタッフがものすごく丁寧で、部屋も広く綺麗でした。
1	2	また、食事は豪華で美味しかったです。

【0035】

文書IDは、個々の文章を識別する情報であり、ここでは重複しない数値である。各文章IDは、テキストデータIDとの関連も保持されている。したがって、一つの文書IDについては、表1に示した属性情報も関連づけられていることになる。

【0036】

一つのテキストデータは、同一人により、ある特定の宿泊施設を利用した際の感想などが表されたものであるが、各文章に着目すると異なる観点について述べていることが多い。表2の例では、スタッフや部屋に関する文章ID「1」や食事に関する文章ID「2」など、異なる観点について宿泊施設に関する感想が述べられている。

【0037】

後述するクラス抽出手段12では、各文章のトピックとなるクラスを抽出するが、もし、仮にテキストデータを分類する場合、テキストデータに異なる観点の文章が複数含まれていると、適切なトピックとはいえない結果となりうる。しかし、本発明では、テキストデータから文章を抽出するので、後述するクラス抽出手段12による抽出精度を向上させることができる。

【0038】

このように、テキストデータから抽出された文章から、第1単語群及び第2単語群を抽出する。第1単語群とは、各文章に含まれる第1品詞に該当する単語からなり*

、第2単語群とは、各文章に含まれる第2品詞に該当する単語からなる。第1品詞及び第2品詞は予め定めておく。

【0039】

単語抽出手段11は、各文章IDで特定される文章を読み込み、公知の形態素解析手法を適用することで、一つの文章の中から第1品詞及び第2品詞に該当する単語の組を抽出する。本実施形態では、第1品詞として名詞、第2品詞として形容詞（形容動詞を含む）とする。例えば、表2の文章IDについては、名詞として「スタッフ」「部屋」が得られ、形容詞として「丁寧」「広い」「綺麗」が得られる。

【0040】

なお、第1品詞及び第2品詞は、名詞や形容詞に限定されない。また、第1品詞及び第2品詞は一つの品詞に限らず、二つの品詞の組み合わせでもよい。例えば、第1品詞として名詞、第2品詞として形容詞又は動詞を採用してもよい。

【0041】

そして、単語抽出手段11は、文章より抽出された第1単語群及び第2単語群から、共起行列を集計する。共起行列とは、第1単語群に属する単語と、第2単語群に属する単語との組み合わせの個数を表したものである。表3に共起行列を例示する。

【0042】

【表3】

		形容詞(第2品詞)				
		良い	広い	美味しい	清潔	...
名詞 (第1品詞)	部屋	1326	1574	125	1168	
	朝食	426	52	1097	69	
	風呂	397	355	44	198	
	対応	795	40	45	118	
	...					

【0043】

名詞（第1品詞）である第1単語群に属する単語「部屋」「朝食」「風呂」「対応」などが行方向に並び、形容詞（第2品詞）である第2単語群に属する単語「良い」「広い」「美味しい」「清潔」などが列方向に並んでいる。各数値は、一つの文章の中に、例えば、「部屋」と「良い」との組み合わせが存在すれば、一つカウント

する。

【0044】

したがって、単語抽出手段11は、文章ごとに、形態素解析により名詞及び形容詞を抽出し、それらの名詞と形容詞との組み合わせを集計して上記の共起行列を作成する。

【0045】

クラス抽出手段12は、前記共起行列を入力とし、第1単語群に属する単語及び第2単語群に属する単語で構成される、文章のトピックとなる複数のクラスを抽出する潜在意味解析法を実行することにより、各クラスを条件とした第1単語群に属する単語の第1条件付確率、及び各クラスを条件とした第2単語群に属する単語の第2条件付確率を求める。

【0046】

潜在意味解析法とは、自然言語処理の技法の一つであり、文書群と文書に含まれる用語群について、それらに関連した概念の集合を生成することで、その関係を分析する手法である。潜在意味解析法の実例としては、LSI (Latent Semantic Indexing)、LDA (Latent Dirichlet Allocation)、PLSA (Probabilistic Latent Semantic Analysis) を挙げることができる。

【0047】

本実施形態では、PLSAを用いて説明する。図2は、PLSAの概念図である。図2(a)に示すように、PLSAは、文書分類に用いられるクラスタリング手法の一つであり、一般には、文章Dと、その文章に含まれる単語Wの間に潜在的なクラスCがあると想定し、文章D及び単語Wの組み合わせで構成されるクラスCを抽出するものである。PLSAによるクラス抽出は、各クラ

*スCに属する文章Dの条件付確率及び各クラスCに属する単語Wの条件付確率及びクラスCの確率がEMアルゴリズムにより計算される。

【0048】

本実施形態では、このようなPLSAに入力するデータは、上述した共起行列である。PLSAは、このような共起行列を入力として、図2(b)に示すように、第1単語群に属する単語N(名詞)と、第2単語群に属する単語A(形容詞)との間に潜在的なクラスCがあると想定し、名詞Nと形容詞Aの組み合わせで構成されるクラスCを抽出するものである。すなわち、クラス抽出手段12は、共起行列を入力としてPLSAを実行することで、各クラスCを条件とした第1単語群に属する単語N(名詞)の第1条件付確率として $P(N|C)$ 、及び各クラスCを条件とした第2単語群に属する単語A(形容詞)の第2条件付確率として $P(A|C)$ を計算する。PLSAの具体的な計算方法は、非特許文献7など、公知の技法を用いて実行することができる。

【0049】

表4に、PLSAにより計算されたクラスに属する名詞及び形容詞を例示する。表4には、複数作成されたクラスのうち、2つのクラスC5とC7に属する名詞及び形容詞が示されている。それぞれ条件付確率が高い順に単語を並べている。

【0050】

【表4】

クラスC5				クラスC7			
$P(N C5)$	名詞	$P(A C5)$	形容詞	$P(N C7)$	名詞	$P(A C7)$	形容詞
30%	部屋	29%	綺麗	26%	朝食	59%	美味しい
8%	ホテル・宿	25%	清潔	5%	バイキング	6%	良い
5%	満足	13%	広い	4%	満足	5%	豊富
4%	風呂	6%	良い	4%	種類	5%	残念
2%	駅	4%	新しい	3%	パン	3%	大変
2%	利用	4%	快適	3%	料理	3%	十分
2%	フロント	2%	静か	3%	食事	2%	嬉しい

【0051】

クラスC5には、所属確率(第1条件付確率)が上位である単語は「部屋」や「ホテル」という名詞であり、所属確率(第2条件付確率)が上位である単語は「綺麗」「清潔」「広い」といった形容詞である。このようなクラスC5に所属する名詞及び形容詞の所属確率に基づいて、クラスC5の意味を解釈することができる。本実施形態では、宿泊施設のクチコミデータを元にクラス抽出しているので、クラスの意味はクチコミデータが話題としている主題(トピック)であるといえる。

【0052】

例えば、クラスC5は、所属確率が上位である単語に基づけば、部屋の綺麗さを表すトピックであると解釈することができる。同様に、クラスC7は、朝食の美味しさを表すトピックであると解釈することができる。

【0053】

仮に、PLSAの入力データとして、従来のように、文章Dと単語Wとの共起行列を用いた場合、表5のようなクラスが抽出される。

【0054】

【表5】

クラスX			
P(D X)	文章D	P(W X)	単語W
1.9%	毎回宿泊させてもらっていますが、快適に過ごすことができます。	35%	宿泊
1.9%	いつもは格安チェーンのホテルに宿泊していたが、ここは以前から気になっていた。	17%	快適
1.7%	浴室も脱衣所も広く、露天風呂もあり快適でした。	9%	凄い
1.7%	土曜宿泊で料金は妥当でしたが、平日利用なら格安です。	8%	安い
1.6%	スタッフの方々がとても親切で、ホテル全体が落ち着いた良い雰囲気でした。	7%	雰囲気

【0055】

例えば、クラスXに属する文章Dの所属確率と、単語Wの所属確率が計算される。これらの文書D及び単語Wの所属確率に基づいた場合、クラスXが何を表しているかを解釈することが難しい場合もある。この例では、所属確率の高い文書では、宿泊の快適性や料金の安さ、さらにスタッフの対応に関する感想を述べており、また所属確率の高い単語からもクラスXがどのようなトピックを意味しているのか解釈しにくい。

【0056】

これは、入力データとなる文章Dと単語Wとの共起行列では、その単語が含まれる文章に対して“1”の値が与えられるが、ほとんどが0の値となる非常に疎な共起行列となり、文章間、単語間で差が出にくいデータとなるため、文章と単語の共通する意味を抽出することが難しくなるといえる。

【0057】

しかしながら、本発明では、文章及び単語からなる共起行列ではなく、文章に含まれる単語同士（名詞及び形容詞）からなる共起行列を作成したため、単語間で出現頻度に差が出やすく、これにPLSAを実行することで*

*、表4のように、抽出されたクラスの意味を解釈しやすくすることができる。

【0058】

PLSAは、クラス数を予め設定する必要があり、また、初期値依存性があるため初期値によって結果が異なる。そこで、本実施形態のクラス抽出手段12では、クラス数として範囲を持たせて複数設定し、初期値を変えてそれぞれのクラス数でPLSAを複数回実行し、それぞれの結果の情報量基準の値を計算する。そして、その全結果の中で情報量基準が最適となる結果を採用する。情報量基準の計算は、公知の方法（例えば「小西貞則、北川源四郎:情報量基準,朝倉書店,2004」参照）により行うことができる。なお、クラス数は、このような情報量基準に基づいて決定する場合に限定されず、任意に定めてもよい。

【0059】

本実施形態では、表6に示すように、クラス抽出手段12により18個のクラスが抽出され、それぞれのクラスの解釈がなされた。

【0060】

【表6】

クラス	意味	クラス	意味
C1	部屋環境	C10	サービス嬉しさ
C2	駅近さ	C11	コストパフォーマンス
C3	値段手頃さ	C12	部屋気持ちよさ
C4	良さ	C13	建物綺麗さ
C5	部屋綺麗さ	C14	スタッフ丁寧さ
C6	チェックイン対応	C15	場所便利さ
C7	朝食美味しさ	C16	朝食多さ
C8	部屋音環境	C17	部屋風呂悪さ
C9	ホテル旅行楽しさ	C18	部屋風呂広さ

【0061】

スコア計算手段13は、第1条件付確率及び第1単語群の発生数、並びに第2条件付確率及び第2単語群の発生数に基づいて、各クラスを条件とした各文章の条件付

確率を計算し、各テキストデータに対する各クラスのスコアを求める。

【0062】

各クラス C_k を条件とした各文章 D_h の条件付確率である $P(D_h | C_k)$ を各文章 D_h に対する各クラス C_k のスコアとして計算する。スコア計算手段13は、 $P(D_h | C_k)$ を次のように計算する。

【0063】

なお、 k は、PLSAで作成されたクラスを特定する*

$$(1) P(Dn_h | C_k) = \sum_i P(Dn_h | N_i) P(N_i | C_k)$$

$$(2) P(Da_h | C_k) = \sum_j P(Da_h | A_j) P(A_j | C_k)$$

$$(3) P(Dn_h | N_i) = \frac{1}{n(N_i)} \quad P(Da_h | A_j) = \frac{1}{n(A_j)}$$

$$(4) P(D_h | C_k) = \frac{1}{2} P(Dn_h | C_k) + \frac{1}{2} P(Da_h | C_k)$$

【0065】

各文章 D_h について、名詞（第1品詞）によって定義される文章を Dn_h 、形容詞（第2品詞）によって定義される文章を Da_h とする。 $P(D_h | C_k)$ を計算するにあたり、 $P(Dn_h | C_k)$ と $P(Da_h | C_k)$ を計算する。これらはそれぞれ式（1）（2）で計算される。単語 W が含まれる文章の数を $n(W)$ とすると、 $P(Dn_h | N_i)$ と $P(Da_h | A_j)$ は式（3）として計算される。すなわち、 $P(Dn_h | N_i)$ は第1単語群の発生数である $n(N_i)$ の逆数、 $P(Da_h | A_j)$ は第2単語群の発生数 $n(A_j)$ の逆数として得られる。第1条件付確率である $P(N_i | C_k)$ と第2条件付確率である $P(A_j | C_k)$ は、PLSAの実行結果によって得られる。

【0066】

$P(Dn_h | C_k)$ と $P(Da_h | C_k)$ は C_k を条件とした文章 D_h において重みは同じといえるので、式（4）により $P(D_h | C_k)$ を計算する。この $P(D_h | C_k)$ を各文章 D_h に対する各クラス C_k のスコアとしてもよいが、この値は、文章の数が多いほど値が小さくなり、この値だけではクラスと文章の関係の強さが※

*番号であり、クラスの総数を最大とする自然数である。

h は、文章を特定する番号（文章ID）であり、文章の総数を最大とする自然数である。

【0064】

【数1】

※分かり難い。

【0067】

このため、上述したスコアである $P(D_h | C_k)$ を、各文章の発生確率である $P(D_h)$ で除した値を用いてもよい。本実施形態では、各文章の発生確率が一様分布に従うと仮定し、 $P(D_h)$ は、文章の総数の逆数とする。

【0068】

このように、 $P(D_h | C_k)$ と $P(D_h)$ との比をもって文章 D_h におけるクラス C_k のスコアとする。この値が1を超えるということは、文章 D_h の発生確率はクラス C_k を条件とすることで上昇し、クラス C_k との関係が強いということである。このように、上述したスコアである $P(D_h | C_k)$ を、一様分布と仮定した各文章の発生確率である $P(D_h)$ で除した値をスコアとすることで、各文章 D_h とクラス C_k の関係の強さを把握しやすくすることができる。表7に各文章 D_h に対する各クラス C_k のスコア $P(D_h | C_k)$ を $P(D_h)$ で除したものを例示する。

【0069】

【表7】

テキスト データID	文章 ID(h)	クラス C_k			
		C1	C2	...	C18
1	1	$P(D1 C1)/P(D1)=3.1$	$P(D1 C2)/P(D1)=0.9$		$P(D1 C18)/P(D1)=1.1$
1	2	$P(D2 C1)/P(D2)=1.4$	$P(D2 C2)/P(D2)=0.2$		$P(D2 C18)/P(D2)=2.4$
2	3	$P(D3 C1)/P(D3)=0.8$	$P(D3 C2)/P(D3)=5.8$		$P(D3 C18)/P(D3)=0.9$
2	4	$P(D4 C1)/P(D4)=1.2$	$P(D4 C2)/P(D4)=3.2$		$P(D4 C18)/P(D4)=1.0$
2	5	$P(D5 C1)/P(D5)=0.6$	$P(D5 C2)/P(D5)=1.8$		$P(D5 C18)/P(D5)=1.6$
...	...	...	...		...

【0070】

さらにこの文章ID単位に計算された各クラスのスコ

アをテキストデータID単位に見たときに、そのテキス

トデータを構成する複数の文章それぞれに対する各クラスのスコアを集約することで、テキストデータIDそれぞれに対する各クラスのスコアを決定する。複数存在する文章単位のスコアをテキストデータ単位に1つに集約する方法としては、最大値や平均値などを計算することが挙げられるが、本実施形態では、クラス毎のスコアの最大値を取ることで、テキストデータIDそれぞれに対*

*する各クラスのスコアを決定する。

【0071】

表8を用いて具体的に説明する。IDが「1」であるテキストデータをテキストデータ「1」と表記し、IDが「1」である文章を文章「1」と表記する。

【0072】

【表8】

テキストデータID	文章ID(h)	クラスCk				
		C1	C2	C3	...	C18
1	1	3.1	0.9	2.0		1.1
1	2	1.4	0.2	5.5		2.4
2	3	0.8	5.8	1.3		0.9
2	4	1.2	3.2	1.7		1.0
2	5	0.6	1.8	2.6		1.6
...	...	...	...			...

【0073】

例えば、テキストデータ「1」は、文章「1」、文章「2」から構成されている。この文章「1」、文章「2」のそれぞれに対する各クラスC1～C18のスコアについて、クラス毎に最大値（文章「1」と文章「2」のうち大きいスコア）を求める。

【0074】

クラスC1については、文書「1」に対するクラスC1のスコアは「3.1」であり、文書「2」に対するクラスC1のスコアは「1.4」である。したがって、「3.1」が最大値となる。この最大値がテキストデータ「1」に対するクラスC1のスコアとなる。以下同様に、クラスC2～C18についてクラス毎に最大値を計算することで、テキストデータ「1」に対する各クラスのスコアとする。このような最大値を求めてテキストデータに対する各クラスのスコアとする計算を、全テキストデータについて実行する。表8の斜体字で表されたスコアがテキストデータに対する各クラスのスコアである。このようにして、各テキストデータに対して、各クラスのスコアを得ることができる。

【0075】

モデル化手段は、各テキストデータの各クラスのスコア及び属性情報を変数として、それらの関係をモデル化するものである。ここでいうモデル化とは、変数間の定*

※量的な関係を定式化した数理モデルを構築することを指し、そのモデルを用いることで一方の変数から他方の変数の状態を計算することができる。このようなモデル化の一例として、本実施形態に係るモデル化手段14では、各テキストデータの各クラスのスコア及び属性情報を確率変数として、ベイジアンネットワークの計算を実行することで、クラスと属性情報の関係をモデル化する。その他のモデル化手法としては、例えば、回帰分析や決定木分析を挙げることができる。

【0076】

各クラスのスコアは連続値であるが、ベイジアンネットワークで扱う変数は質的変数となるので、適当な閾値を設定するなどして離散化する。ここでは、各テキストデータの各クラスのスコアは、例えば、スコアが3を超えればHigh、3以下であればLowという2値を取る離散的な確率変数とする。この閾値は、各文章の内容とそのスコアの大きさを目視することで決定する。

【0077】

ここでは、表9に表8のスコアを離散的な確率変数とした結果を示す。すなわち、テキストデータに対するクラスのスコア（表8の斜体字のスコア）について離散的な確率変数とした。

【0078】

【表9】

テキストデータID	クラスCk				
	C1	C2	C3	...	C18
1	High	Low	High		Low
2	Low	High	Low		Low
...	...	...	...		...

【0079】

また、属性情報に関しても離散的な確率変数となるように前処理しておく。表10に、離散化した属性情報の

取り得る値を例示する。

【0080】

【表10】

	属性情報								
	利用者属性		施設属性		宿泊内容		項目得点		
確率変数	性別	年代	ラウンジ	LAN	宿泊料金	部屋サイズ	総合得点	朝食得点	サービス得点
取り得る値	男	10代	あり	あり	～10000	～20m ²	3点以下	3点以下	3点以下
	女	20代	なし	なし	10001～15000	～30m ²	4点	4点	4点
		30代			15001～20000	～40m ²	5点	5点	5点
		40代							
		50代							
		60代							
		70代							

【0081】

そして、各テキストデータに関連づけられた属性情報のうち連続的な値については、表10に示したように離散的な値に置き換える。例えば、表1に示した属性情報*

*は表11のように一部が置き換えられる。なお、項目得点については一部省略している。

【0082】

【表11】

テキストデータID	属性情報								
	利用者属性		施設属性		宿泊内容		項目得点		
	性別	年代	ラウンジ	LAN	宿泊料金	部屋サイズ	総合得点	朝食得点	サービス得点
1	男	20代	なし	あり	10001～15000	～30m ²	4点	3点以下	5点
2	女	30代	あり	あり	～10000	～20m ²	3点以下	4点	3点以下
3	男	50代	あり	あり	10001～15000	～30m ²	5点	4点	3点以下

【0083】

このような前処理を行った結果、モデル化手段14においては、一つのテキストデータIDについて、クラスごとのスコアと、属性情報とが関連づけられた表12の※30

※ようなデータを入力とする。

【0084】

【表12】

テキストデータID	クラスのスコア					属性情報								
	C1	C2	C3	...	C18	利用者属性		施設属性		宿泊内容		項目得点		
						性別	年代	ラウンジ	LAN	宿泊料金	部屋サイズ	総合得点	朝食得点	サービス得点
1	High	Low	High		Low	男	20代	なし	あり	10001～15000	～30m ²	4点	3点以下	5点
2	Low	High	Low		Low	女	30代	あり	あり	～10000	～20m ²	3点以下	4点	3点以下
3	Low	Low	Low		Low	男	50代	あり	あり	10001～15000	～30m ²	5点	4点	3点以下

【0085】

モデル化手段14は、表12に示すような各テキストデータIDのクラスのスコア、及びこれに関連づけられた属性情報とを確率変数として、ベイジアンネットワークの計算を実行することで、クラスと属性情報の関係をモデル化する。ベイジアンネットワークの具体的な計算方法は、非特許文献6など公知の方法により計算するこ

とができる。

【0086】

このベイジアンネットワークにより、上述した複数の確率変数の間の依存関係をグラフ構造によって表わし、その変数間の定量的な関係を条件付き確率によって表わした確率モデルが得られる。すなわち、テキストデータ

のクラス（トピック）と属性情報との確率的関係がモデル化される。

【0087】

図3は、モデル化手段14により得られたベイジアンネットワークの一部である。同図には、クラス抽出手段12により抽出されたクラス（トピック）と、属性情報のうち項目得点とからなるベイジアンネットワークの一部が示されている。ベイジアンネットワークの各ノードは確率変数を表し、リンクはノード間の依存関係を表している。そして、ノード間の依存関係が条件付確率で定

【0088】

例えば、クラスC13（「建物綺麗さ」というトピック）は、清潔得点や風呂得点、部屋得点に依存関係があり、 $P(\text{清潔得点} | C13)$ や $P(\text{風呂得点} | C13)$ 、 $P(\text{部屋得点} | C13)$ というように条件付確率が計算されている。

【0089】

従来では、テキストデータから単語を抽出し、その単語を確率変数としてベイジアンネットワークによるモデルを構築したものがあるが、この場合、単語の数（ノード数）や、単語間の依存関係（リンク数）が多すぎて、モデルを把握したり、解釈することが非常に困難であった。

【0090】

しかしながら、本発明によれば、ベイジアンネットワ*

クラス		総合得点=5点
C5「部屋綺麗さ」	Low	0.31
	High	0.42
C7「朝食美味しさ」	Low	0.33
	High	0.42
C14「スタッフ丁寧さ」	Low	0.32
	High	0.44

【0095】

このような結果からは、例えば、クラスC5「部屋綺麗さ」が「Low」であれば、総合得点が「5点」となる確率が0.31であるのに対し、「部屋綺麗さ」が「High」になると、総合得点が「5点」となる確率は0.42であるということが推論できる。

【0096】

または、推論手段15は、逆に、属性情報の確率変数が取り得る値を変化させた際に、クラスの確率がどの程※

部屋得点	クラスC5「部屋綺麗さ」=High
5点	0.47
3点以下	0.23

【0099】

このような結果からは、部屋得点が「3点以下」であれば、「部屋綺麗さ」が「High」となる確率が0.23であるのに対し、部屋得点が「5点」になると、「

*ークでは単語そのものを適用対象とせず、PLSAにより抽出されたクラスを対象とする。これにより、単語を適用対象とするよりも、ベイジアンネットワークによるモデルがシンプルとなり、モデルの把握や解釈を容易とすることができる。

【0091】

推論手段15は、クラスの確率変数を変化させたときの属性情報の状態、又は属性情報の確率変数を変化させた時のクラスの状態を推論する。

【0092】

具体的には、推論手段15は、クラスの確率変数及び属性情報の確率変数の関係を表す条件付確率に基づいて、クラスの確率変数が取り得る値を変化させた際に、属性情報の確率がどの程度変化するかを計算する。

【0093】

例えば、クラスC5、C7、C14（それぞれトピックは、「部屋綺麗さ」「朝食美味しさ」「スタッフ丁寧さ」）の確率変数は取り得る値は「High」「Low」である。そして、総合得点の取り得る値は「3点以下」「4点」「5点」である。この各クラスの値が「Low」から「High」に変化したとき、総合得点が「5点」である確率がどの程度変化するかを計算する。この結果を表13に示す。

【0094】

【表13】

※度変化するかを計算する。

【0097】

例えば、部屋得点の確率変数が「3点以下」から「5点」に変化したとき、クラスC5「部屋綺麗さ」が「High」である確率がどの程度変化するかを計算する。この結果を表14に示す。

【0098】

【表14】

部屋綺麗さ」が「High」となる確率は0.47であるということが推論できる。

【0100】

このような推論結果は、業務改善すべき点を効率的に

見いだすことに役立てることができる。具体的には、表13のような推論結果によれば、どの様な観点（トピック）が総合得点をどの程度押し上げる、又は押し下げるかを定量的に把握することができる。したがって、どの観点（トピック）から業務改善やサービスの充実を図ればよいか、優先順位を決定することができる。例えば、総合得点が5点となる確率が最も高いのは、クラスC14「スタッフ丁寧さ」であるから、「スタッフ丁寧さ」のスコアが向上するような業務改善等を優先的に行う、などと意思決定することができる。

【0101】

次に、本実施形態に係る分析装置1の動作について説明する。図4は、分析装置での処理を示すフローチャートである。

【0102】

まず、テキストデータから共起行列を作成する（ステップS1：単語抽出ステップ）。具体的には、単語抽出手段11が、テキストデータから文章を抽出し、各文章から、予め定めた第1品詞及び第2品詞のそれぞれに該当する第1単語群及び第2単語群を抽出し、各文章に含まれている第1単語群に属する単語及び第2単語群に属する単語の組み合わせの個数を表す共起行列を作成する。具体例については、上述したので説明は省略する。

【0103】

次に、共起行列を入力として潜在意味解析法を実行する（ステップS2：クラス抽出ステップ）。具体的には、クラス抽出手段12が共起行列を入力とし、第1単語群に属する単語及び第2単語群に属する単語で構成される複数のクラスを抽出する潜在意味解析法を実行する。これにより、各クラスを条件とした第1単語群に属する単語の第1条件付確率、及び各クラスを条件とした第2単語群に属する単語の第2条件付確率が得られる。具体例については、上述したので説明は省略する。

【0104】

次に、各テキストデータに対する各クラスのスコアを計算する（ステップS3：スコア計算ステップ）。具体的には、スコア計算手段13が、第1条件付確率及び第1単語群の発生数、並びに第2条件付確率及び第2単語群の発生数に基づいて、各クラスを条件とした各文章の条件付確率を各文章に対する各クラスのスコアとして求め、それをテキストデータ単位に集約することで、各テキストデータに対する各クラスのスコアを求める。具体例については上述したので説明は省略する。

【0105】

次に、クラスと属性情報の関係をモデル化する（ステップS4：モデル化ステップ）。具体的には、モデル化手段14が、クラス及び属性情報を確率変数として、ベイジアンネットワークを計算することで、クラスと属性情報の関係をモデル化する。具体例については上述したので説明は省略する。

【0106】

次に、クラスの確率変数を変化させたときの属性情報の状態、又は属性情報の確率変数を変化させた時のクラスの状態を推論する（ステップS5：推論ステップ）。具体的には推論手段15が、クラスの確率変数及び属性情報の確率変数の関係を表す条件付確率に基づいて、クラスの確率変数が取り得る値を変化させた際に、属性情報の確率がどの程度変化するかを出力する。具体例については上述したので説明は省略する。

10 【0107】

以上に述べたように、本実施形態に係る分析方法、分析装置及び分析プログラムによれば、テキストデータからクラス（トピック）を抽出し、各クラスと属性情報との関係をベイジアンネットワークによりモデル化した。そして、そのモデルを用いて、クラスが変化したときの属性情報の変化や、属性情報が変化したときのクラスの変化を定量的に得ることができる。これにより、業務改善すべき点や利用客のニーズを抽出することができる。また、業務改善を施したときに、その結果がどのように変化するのかシミュレーションすることができる。

【0108】

また、ベイジアンネットワークによるモデルの構築は、テキストデータから文章を抽出し、さらに文章から抽出した単語を確率変数として用いるのではなく、PLSAにより得られたクラス（トピック）を確率変数として用いた。これにより、単語を適用対象とするよりも、ベイジアンネットワークによるモデルがシンプルとなり、モデルの把握や解釈を容易とすることができる。

【0109】

30 さらに、本発明では、文章及び単語からなる共起行列ではなく、文章に含まれる単語同士（名詞及び形容詞）からなる共起行列を用いてPLSAを実行したので、クラスの意味の解釈を容易に行うことができる。

【0110】

また、このように抽出されたクラスに対する各文章のスコアを計算する本発明の方法により、そのクラスをベイジアンネットワークを用いたモデル化における変数として処理することができる。

【0111】

40 このように、本発明に係る分析方法によれば、テキストデータに含まれる文章についてクラスの意味付け（トピックの決定）を容易とすることができ、当該トピックに基づいてベイジアンネットワークによるモデル化をすることで、モデルが複雑になることを回避し、さらに、そのモデル化結果において、条件を変化させたときにどのような結果となりうるのかを推論することができる。

【0112】

50 なお、本発明を上述した実施形態に基づいて説明したが、本発明は上記実施形態に限定されない。例えば、本発明は、上記の実施形態で説明したフローチャートの処

29

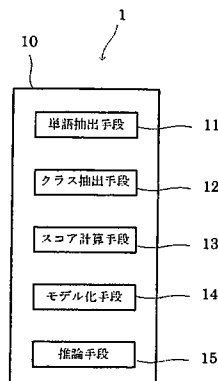
理手順が開示する分析方法であるとしてもよい。また、一台の分析装置1において各手段11～15による処理を実行させたが、このような態様に限らず、複数の分析装置にて各手段を分散して実行させてもよい。

【符号の説明】

【0113】

1 分析装置

【図1】

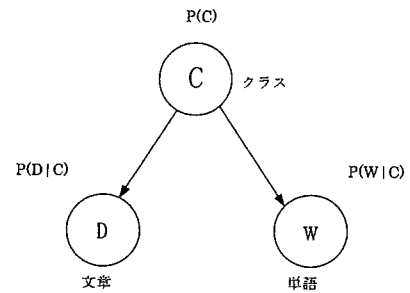


30

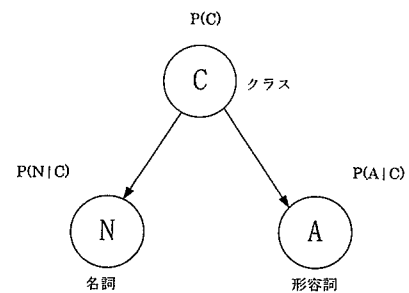
- 10 分析プログラム
- 11 単語抽出手段
- 12 クラス抽出手段
- 13 スコア計算手段
- 14 モデル化手段
- 15 推論手段

【図2】

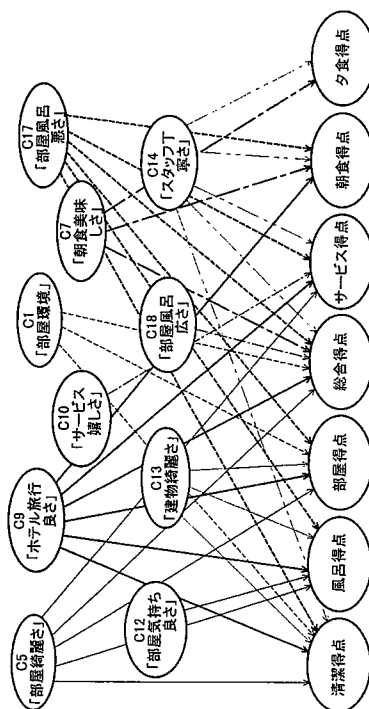
(a)



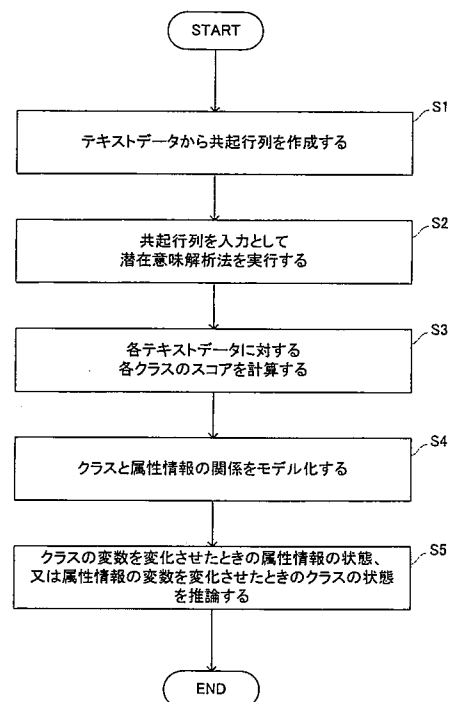
(b)



【図3】



【図4】



フロントページの続き

特許法第30条第2項適用 掲載年月日 平成26年4月23日、 掲載アドレス <http://ja.serviceology.org/events/domestic.html> <http://ja.serviceology.org/publish/domestic2014/index.html> http://ja.serviceology.org/publish/domestic2014_for_participants/proceeding-FullVersion.pdf 集会名 2014年度 サービス学会 第2回国内大会、開催日 平成26年4月29日

特許法第30条第2項適用 掲載年月日 平成26年5月1日、 掲載アドレス <http://www.ai-gakkai.or.jp/jsai2014/proceedings> 集会名 2014年度人工知能学会全国大会(第28回)、開催日 平成26年5月12日

早期審査対象出願

審査官 早川 学

(56)参考文献 米国特許出願公開第2004/0088308 (US, A1)

(58)調査した分野(Int.Cl., DB名)

G06F 17/30

G06F 17/27

G06Q 10/00-99/00