(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 I		
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

1

(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、

2

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。
【請求項2】
請求項1に記載する分析方法において、

3

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。

【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

4

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、これによっ

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.:Probabilisticlatent 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的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。

【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。

【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。

【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】

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、テキストデータを一つずつ読み込み、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。

15

【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。潜在意味解析法の具体例としては、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＊

16

＊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】

19

各クラスＣｋを条件とした各文章Ｄｈの条件付確率であるＰ（Ｄｈ｜Ｃｋ）を各文章Ｄｈに対する各クラスＣｋのスコアとして計算する。スコア計算手段１３は、Ｐ（Ｄｈ｜Ｃｋ）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＊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。

20

【００６４】

【数１】

$$(1)\quad P(Dn_h \mid C_k) = \sum_i P(Dn_h \mid N_i) P(N_i \mid C_k)$$

$$(2)\quad P(Da_h \mid C_k) = \sum_j P(Da_h \mid A_j) P(A_j \mid C_k)$$

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

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

【００６５】

各文章Ｄｈについて、名詞（第１品詞）によって定義される文章をＤｎｈ、形容詞（第２品詞）によって定義される文章をＤａｈとする。Ｐ（Ｄｈ｜Ｃｋ）を計算するにあたり、Ｐ（Ｄｎｈ｜Ｃｋ）とＰ（Ｄａｈ｜Ｃｋ）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（Ｄｎｈ｜Ｎｉ）とＰ（Ｄａｈ｜Ａｊ）は式（３）として計算される。すなわち、Ｐ（Ｄｎｈ｜Ｎｉ）は第１単語群の発生数であるｎ（Ｎｉ）の逆数、Ｐ（Ｄａｈ｜Ａｊ）は第２単語群の発生数ｎ（Ａｊ）の逆数として得られる。第１条件付確率であるＰ（Ｎｉ｜Ｃｋ）と第２条件付確率であるＰ（Ａｊ｜Ｃｋ）は、ＰＬＳＡの実行結果によって得られる。

【００６６】

Ｐ（Ｄｎｈ｜Ｃｋ）とＰ（Ｄａｈ｜Ｃｋ）はＣｋを条件とした文章Ｄｈにおいて重みは同じといえるので、式（４）によりＰ（Ｄｈ｜Ｃｋ）を計算する。このＰ（Ｄｈ｜Ｃｋ）を各文章Ｄｈに対する各クラスＣｋ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※分かり難い。

【００６７】

このため、上述したスコアであるＰ（Ｄｈ｜Ｃｋ）を、各文章の発生確率であるＰ（Ｄｈ）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（Ｄｈ）は、文章の総数の逆数とする。

【００６８】

このように、Ｐ（Ｄｈ｜Ｃｋ）とＰ（Ｄｈ）との比をもって文章Ｄｈにおけるクラスの強さとする。この値が１を超えるということは、文章Ｄｈの発生確率はクラスＣｋを条件とすることで上昇し、クラスＣｋ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（Ｄｈ｜Ｃｋ）を、一様分布と仮定した各文章の発生確率であるＰ（Ｄｈ）で除した値をスコアとすることで、各文章Ｄｈとクラス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Ｄｈに対する各クラスＣｋのスコアＰ（Ｄｈ｜Ｃｋ）をＰ（Ｄｈ）で除したものを例示する。

【００６９】

【表７】

		クラスCk									
テキストデータID	文章ID(h)	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※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								
						利用者属性		施設属性		宿泊内容		項目得点		
	C1	C2	C3	・・・	C18	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量化されている。

【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（清潔得点｜Ｃ１３）やＰ（風呂得点｜Ｃ１３）、Ｐ（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＊

＊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】

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※

※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、「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。

【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】

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】

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】

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

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

１０　　分析プログラム

１１　　単語抽出手段

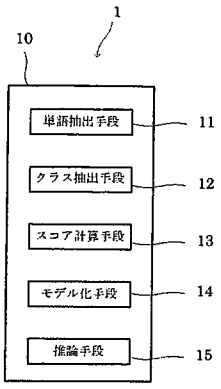
１２　　クラス抽出手段

１３　　スコア計算手段

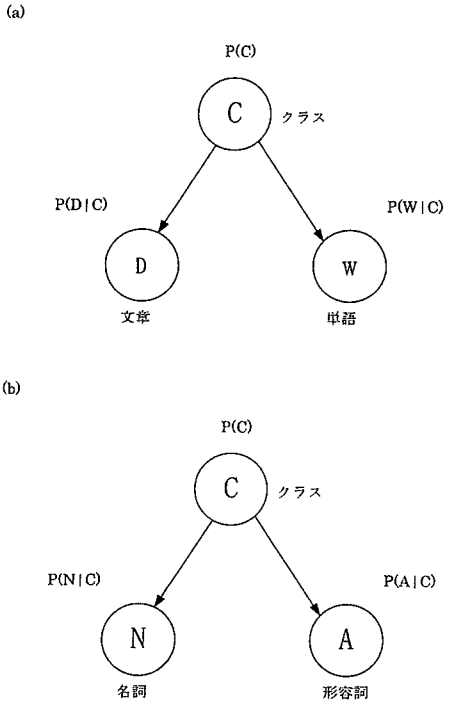
１４　　モデル化手段

１５　　推論手段

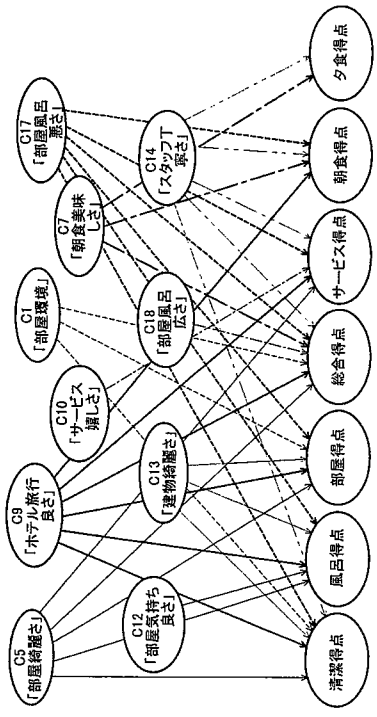
【図１】



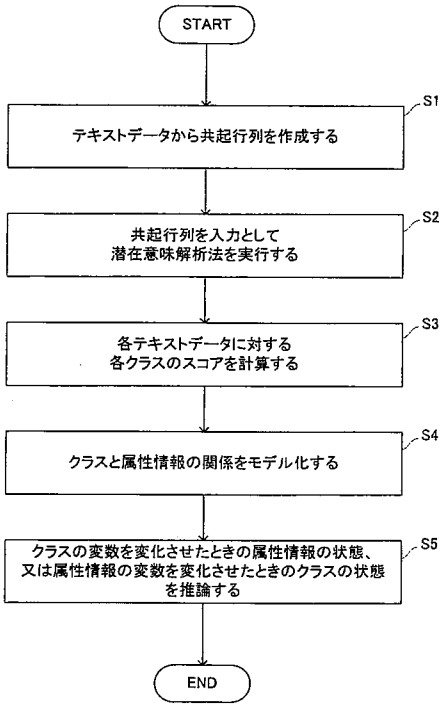
【図２】



【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