

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7221527号  
(P7221527)

(45)発行日 令和5年2月14日(2023.2.14)

(24)登録日 令和5年2月6日(2023.2.6)

(51)Int. Cl.		F I
G 0 6 F 16/30	(2019.01)	G 0 6 F 16/30
G 0 6 F 16/383	(2019.01)	G 0 6 F 16/383
G 0 6 F 40/20	(2020.01)	G 0 6 F 40/20
G 0 6 F 40/216	(2020.01)	G 0 6 F 40/216
G 0 6 F 40/279	(2020.01)	G 0 6 F 40/279

請求項の数 5 (全 25 頁)

(21)出願番号	特願2019-84332(P2019-84332)
(22)出願日	平成31年4月25日(2019.4.25)
(65)公開番号	特開2020-181390(P2020-181390A)
(43)公開日	令和2年11月5日(2020.11.5)
審査請求日	令和4年1月20日(2022.1.20)

(73)特許権者	517219410 株式会社アナリティクスデザインラボ 東京都中野区東中野1-58-8 パーク ハビオ東中野204
(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 東京都中野区東中野一丁目58番8号 パ ークハビオ東中野204 株式会社アナリ ティクスデザインラボ内

審査官 和田 財太

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】

分析装置が実行するテキストデータの分析方法であって、

前記テキストデータに含まれている第1語群に属する語及び第2語群に属する語の組み合わせの頻度に基づく要素からなる共起行列を作成する共起行列作成ステップと、

前記共起行列を入力とし、第1語群に属する語及び第2語群に属する語で構成される複数のトピックを抽出する潜在意味解析法を実行することにより、各トピックを条件とした第1語群に属する語の第1条件付確率、及び各トピックを条件とした第2語群に属する語の第2条件付確率を求めるトピック抽出ステップと、

前記第1条件付確率及び第1語群の出現頻度、並びに前記第2条件付確率及び第2語群の出現頻度に基づいて、各トピックを条件とした各前記テキストデータの条件付確率を計算し、前記条件付確率に基づいて各前記テキストデータに対する各トピックのスコアを求めるスコア計算ステップと、を備え、

前記共起行列作成ステップは、

前記テキストデータから前記第1語群に属する語及び前記第2語群に属する語の組み合わせの頻度を要素とする実測共起行列を作成し、

前記テキストデータから前記第1語群に属する語及び前記第2語群に属する語の組み合わせの期待頻度を要素とする期待共起行列を作成し、

前記期待共起行列の各要素に対する前記実測共起行列の各要素の差分あるいは比率を各要素とする前記共起行列を作成する

ことを特徴とする分析方法。

**【請求項 2】**

請求項 1 に記載の分析方法であって、  
前記共起行列作成ステップは、  
前記テキストデータから文章を抽出し、各文章に含まれている前記第 1 語群に属する語及び前記第 2 語群に属する語の組み合わせの頻度を要素とする前記実測共起行列を作成し

、  
前記テキストデータから文章を抽出し、各文章に含まれている前記第 1 語群に属する語及び前記第 2 語群に属する語の組み合わせの期待頻度を要素とする前記期待共起行列を作成し、

前記期待共起行列の各要素に対する前記実測共起行列の各要素の差分あるいは比率を各要素とする前記共起行列を作成し、

前記スコア計算ステップは、前記第 1 条件付確率及び第 1 語群の出現頻度、並びに前記第 2 条件付確率及び第 2 語群の出現頻度に基づいて、各トピックを条件とした各文章の条件付確率を計算し、前記条件付確率に基づいて各前記テキストデータに対する各トピックのスコアを求める

ことを特徴とする分析方法。

**【請求項 3】**

請求項 1 に記載する分析方法であって、  
前記テキストデータは、カテゴリに分類されたテキスト部を含み、  
前記共起行列作成ステップは、

第 1 のカテゴリに分類された前記テキスト部から抽出した前記第 1 語群に属する語、及び第 2 のカテゴリに分類された前記テキスト部から抽出した前記第 2 語群に属する語の組み合わせの頻度を要素とする前記実測共起行列を作成し、

第 1 のカテゴリに分類された前記テキスト部から抽出した前記第 1 語群に属する語、及び第 2 のカテゴリに分類された前記テキスト部から抽出した前記第 2 語群に属する語の組み合わせの期待頻度を要素とする前記期待共起行列を作成し、

前記期待共起行列の各要素に対する前記実測共起行列の各要素の差分あるいは比率を各要素とする前記共起行列を作成する

ことを特徴とする分析方法。

**【請求項 4】**

テキストデータの分析装置であって、  
前記テキストデータに含まれている第 1 語群に属する語及び第 2 語群に属する語の組み合わせの頻度に基づく要素からなる共起行列を作成する共起行列作成手段と、

前記共起行列を入力とし、第 1 語群に属する語及び第 2 語群に属する語で構成される複数のトピックを抽出する潜在意味解析法を実行することにより、各トピックを条件とした第 1 語群に属する語の第 1 条件付確率、及び各トピックを条件とした第 2 語群に属する語の第 2 条件付確率を求めるトピック抽出手段と、

前記第 1 条件付確率及び第 1 語群の出現頻度、並びに前記第 2 条件付確率及び第 2 語群の出現頻度に基づいて、各トピックを条件とした各前記テキストデータの条件付確率を計算し、前記条件付確率に基づいて各前記テキストデータに対する各トピックのスコアを求めるスコア計算手段と、を備え、

前記共起行列作成手段は、  
前記テキストデータから前記第 1 語群に属する語及び前記第 2 語群に属する語の組み合わせの頻度を要素とする実測共起行列を作成し、

前記テキストデータから前記第 1 語群に属する語及び前記第 2 語群に属する語の組み合わせの期待頻度を要素とする期待共起行列を作成し、

前記期待共起行列の各要素に対する前記実測共起行列の各要素の差分あるいは比率を各要素とする前記共起行列を作成する

ことを特徴とする分析装置。

**【請求項 5】**

テキストデータ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に含まれている第 1 語群に属する語及び第 2 語群に属する語の組み合わせの頻度に基づく要素からなる共起行列を作成する共起行列作成手段と、  
 前記共起行列を入力とし、第 1 語群に属する語及び第 2 語群に属する語で構成される複数のトピックを抽出する潜在意味解析法を実行することにより、各トピックを条件とした第 1 語群に属する語の第 1 条件付確率、及び各トピックを条件とした第 2 語群に属する語の第 2 条件付確率を求めるトピック抽出手段と、  
 前記第 1 条件付確率及び第 1 語群の出現頻度、並びに前記第 2 条件付確率及び第 2 語群の出現頻度に基づいて、各トピックを条件とした各前記テキストデータの条件付確率を計算し、前記条件付確率に基づいて各前記テキストデータに対する各トピックのスコアを求めるスコア計算手段として機能させ、  
 前記共起行列作成手段は、  
 前記テキストデータから前記第 1 語群に属する語及び前記第 2 語群に属する語の組み合わせの頻度を要素とする実測共起行列を作成し、  
 前記テキストデータから前記第 1 語群に属する語及び前記第 2 語群に属する語の組み合わせの期待頻度を要素とする期待共起行列を作成し、  
 前記期待共起行列の各要素に対する前記実測共起行列の各要素の差分あるいは比率を各要素とする前記共起行列を作成することを特徴とする分析プログラム。

10

20

**【発明の詳細な説明】****【技術分野】****【0001】**

本発明は、テキストデータから個性的なトピックを抽出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に関する。

**【背景技術】****【0002】**

昨今では、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。

30

**【0003】**

本発明者は、テキストデータから、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知られる PLSA を応用した分析方法を発明した（特許文献 1 参照）。PLSA は、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。

**【0004】**

このような分析方法においても、テキストデータからマイニングを行い、潜在的なトピックを抽出することはできる。しかしながら、PLSA は、元々のテキストデータに高い頻度で発生する単語を元にトピックを抽出する傾向にあり、得られたトピックは典型的で目新しいものではない場合がある。

40

**【先行技術文献】****【特許文献】****【0005】**

【特許文献 1】特開 2016-051220 号公報

**【発明の概要】****【発明が解決しようとする課題】****【0006】**

本発明は、上記事情に鑑みてなされたものであり、テキストデータに低い頻度で発生するような単語であっても、当該単語に基づく個性的なトピックを抽出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】

【0007】

上記課題を解決する本発明の第1の態様は、分析装置が実行するテキストデータの分析方法であって、前記テキストデータに含まれている第1語群に属する語及び第2語群に属する語の組み合わせの頻度に基づく要素からなる共起行列を作成する共起行列作成ステップと、前記共起行列を入力とし、第1語群に属する語及び第2語群に属する語で構成される複数のトピックを抽出する潜在意味解析法を実行することにより、各トピックを条件とした第1語群に属する語の第1条件付確率、及び各トピックを条件とした第2語群に属する語の第2条件付確率を求めるトピック抽出ステップと、前記第1条件付確率及び第1語群の出現頻度、並びに前記第2条件付確率及び第2語群の出現頻度に基づいて、各トピックを条件とした各前記テキストデータの条件付確率を計算し、前記条件付確率に基づいて各前記テキストデータに対する各トピックのスコアを求めるスコア計算ステップと、を備え、前記共起行列作成ステップは、前記テキストデータから前記第1語群に属する語及び前記第2語群に属する語の組み合わせの頻度を要素とする実測共起行列を作成し、前記テキストデータから前記第1語群に属する語及び前記第2語群に属する語の組み合わせの期待頻度を要素とする期待共起行列を作成し、前記期待共起行列の各要素に対する前記実測共起行列の各要素の差分あるいは比率を各要素とする前記共起行列を作成することを特徴とする分析方法にある。

10

【0008】

本発明の第2の態様は、第1の態様に記載の分析方法であって、前記共起行列作成ステップは、前記テキストデータから文章を抽出し、各文章に含まれている前記第1語群に属する語及び前記第2語群に属する語の組み合わせの頻度を要素とする前記実測共起行列を作成し、前記テキストデータから文章を抽出し、各文章に含まれている前記第1語群に属する語及び前記第2語群に属する語の組み合わせの期待頻度を要素とする前記期待共起行列を作成し、前記期待共起行列の各要素に対する前記実測共起行列の各要素の差分あるいは比率を各要素とする前記共起行列を作成し、前記スコア計算ステップは、前記第1条件付確率及び第1語群の出現頻度、並びに前記第2条件付確率及び第2語群の出現頻度に基づいて、各トピックを条件とした各文章の条件付確率を計算し、前記条件付確率に基づいて各前記テキストデータに対する各トピックのスコアを求めることを特徴とする分析方法にある。

20

30

【0009】

本発明の第3の態様は、第1の態様に記載の分析方法であって、前記テキストデータは、カテゴリに分類されたテキスト部を含み、前記共起行列作成ステップは、第1のカテゴリに分類された前記テキスト部から抽出した前記第1語群に属する語、及び第2のカテゴリに分類された前記テキスト部から抽出した前記第2語群に属する語の組み合わせの頻度を要素とする前記実測共起行列を作成し、第1のカテゴリに分類された前記テキスト部から抽出した前記第1語群に属する語、及び第2のカテゴリに分類された前記テキスト部から抽出した前記第2語群に属する語の組み合わせの期待頻度を要素とする前記期待共起行列を作成し、前記期待共起行列の各要素に対する前記実測共起行列の各要素の差分あるいは比率を各要素とする前記共起行列を作成することを特徴とする分析方法にある。

40

【0010】

本発明の第4の態様は、テキストデータの分析装置であって、前記テキストデータに含まれている第1語群に属する語及び第2語群に属する語の組み合わせの頻度に基づく要素からなる共起行列を作成する共起行列作成手段と、前記共起行列を入力とし、第1語群に属する語及び第2語群に属する語で構成される複数のトピックを抽出する潜在意味解析法を実行することにより、各トピックを条件とした第1語群に属する語の第1条件付確率、及び各トピックを条件とした第2語群に属する語の第2条件付確率を求めるトピック抽出手段と、前記第1条件付確率及び第1語群の出現頻度、並びに前記第2条件付確率及び第2語群の出現頻度に基づいて、各トピックを条件とした各前記テキストデータの条件付確

50

率を計算し、前記条件付確率に基づいて各前記テキストデータに対する各トピックのスコアを求めるスコア計算手段と、を備え、前記共起行列作成手段は、前記テキストデータから前記第1語群に属する語及び前記第2語群に属する語の組み合わせの頻度を要素とする実測共起行列を作成し、前記テキストデータから前記第1語群に属する語及び前記第2語群に属する語の組み合わせの期待頻度を要素とする期待共起行列を作成し、前記期待共起行列の各要素に対する前記実測共起行列の各要素の差分あるいは比率を各要素とする前記共起行列を作成することを特徴とする分析装置にある。

#### 【0011】

本発明の第5の態様は、テキストデータ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に含まれている第1語群に属する語及び第2語群に属する語の組み合わせの頻度に基づく要素からなる共起行列を作成する共起行列作成手段と、前記共起行列を入力とし、第1語群に属する語及び第2語群に属する語で構成される複数のトピックを抽出する潜在意味解析法を実行することにより、各トピックを条件とした第1語群に属する語の第1条件付確率、及び各トピックを条件とした第2語群に属する語の第2条件付確率を求めるトピック抽出手段と、前記第1条件付確率及び第1語群の出現頻度、並びに前記第2条件付確率及び第2語群の出現頻度に基づいて、各トピックを条件とした各前記テキストデータの条件付確率を計算し、前記条件付確率に基づいて各前記テキストデータに対する各トピックのスコアを求めるスコア計算手段として機能させ、前記共起行列作成手段は、前記テキストデータから前記第1語群に属する語及び前記第2語群に属する語の組み合わせの頻度を要素とする実測共起行列を作成し、前記テキストデータから前記第1語群に属する語及び前記第2語群に属する語の組み合わせの期待頻度を要素とする期待共起行列を作成し、前記期待共起行列の各要素に対する前記実測共起行列の各要素の差分あるいは比率を各要素とする前記共起行列を作成することを特徴とする分析プログラムにある。

#### 【発明の効果】

#### 【0012】

本発明によれば、テキストデータに低い頻度で発生するような単語であっても、当該単語に基づく個性的なトピックを抽出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】

#### 【0013】

【図1】本実施形態に係る分析方法を実装した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】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】

#### 【0014】

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。

#### 【0015】

〈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。

#### 【0016】

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。

10

20

30

40

50

## 【0017】

このような分析プログラムは、テキストデータを処理対象とする。テキストデータとは、文章を符号化したデータである。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。テキストデータの符号化の方式（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などの文字コードで表現されている。

## 【0018】

本実施形態では、テキストデータとして、日本の特許出願に添付された要約書の文章を用いる。具体的には、要約書及び特許請求の範囲に「電気」及び「車」を含む10年分（出願日が2007年1月1日から2016年12月31日）の電気自動車に関する特許出願（26,419件）を抽出し、その特許出願の要約書の記載をテキストデータとする。

## 【0019】

## 【表1】

テキストデータID	文章ID	要約文
1	1	電気自動車の充電完了後における無駄な電力消費を抑制する。
1	2	制御部は、計測部の計測結果から電気自動車の充電完了を判断し、開閉部を制御して給電路を開成させる。
1	3	そして、開閉部によって給電路が開成されると、電気自動車と商用交流電源とが完全に切り離されるため、充電スタンドから電気自動車に待機電流が流れたり、電気自動車の蓄電池から充電スタンドの方へ自然放電されることがなくなる。
1	4	その結果、電気自動車の充電完了後における無駄な電力消費を抑制することができる。
2	5	電気自動車に電気を供給するための充電式バッテリーは、パイプラインに沿った輸送に適合するカプセルの中に封入される、少なくとも一つの着脱交換式再充電可能電池を含む。
3	6	非接触充電の送電効率の向上を図ることができる非接触充電器の車輛への取付構造を提供する。
3	7	受電コイルを有する受電ユニットの電気自動車への取付構造において、受電ユニットを電気自動車に有する一対の後輪の間に配置する。
4	8	給電スタンドと車両の間の電波環境の診断を行うことができる非接触充電システムを提供する。
4	9	充電システムは、給電スタンドと電気自動車に共に有線接続される診断装置を備えている。
4	10	給電スタンドと電気自動車は、無線通信手段と、有線通信手段としても機能する制御手段をそれぞれ備えている。
4	11	診断装置は、有線通信手段としても機能する制御手段を備えている。
4	12	診断装置は、給電スタンドおよび電気自動車との間で有線通信によってそれぞれ診断制御信号をやり取りすることによって、給電スタンドと電気自動車との間の電波環境の診断を行う。

## 【0020】

表1にテキストデータの一例を示す。表1には、4つのテキストデータが例示されている。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。テキストデータは、発明の要約文である。文章IDは、テキストデータに含まれる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。以後、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。

## 【0021】

テキストデータを分析対象とする分析装置1は、共起行列作成手段11、トピック抽出手段12、及びスコア計算手段13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。分析プログラム10は、分析装置1を各手段11～13として機能させるプログラムである。

【0022】

共起行列作成手段11は、テキストデータから共起行列を作成する。共起行列とは、第1語群に属する語、及び第2語群に属する語の組み合わせの頻度に基づく要素からなる行列であり、具体的には、以下のように、実測共起行列と期待共起行列とから作成される。

【0023】

実測共起行列とは、第1語群に属する語及び第2語群に属する語の組み合わせ（共起ペアと称する）を含むテキストデータの頻度（件数）を要素とする行列である。実測共起行列は、次のようにして作成される。

【0024】

まず、共起行列作成手段11は、テキストデータから文章を抽出する。具体的には、共起行列作成手段11は、テキストデータの一つずつ読み込み、各テキストデータについて、句点など一文の末尾に用いられる文字を基準として文章を出力する。例えば、テキストデータID「1」については、表1に示すように4つの文章が抽出される。

【0025】

一つのテキストデータは、発明に関する記載が含まれているが、各文章に着目すると異なる観点で記載されていることが多い。表1のテキストデータID「1」からは、電気自動車の課題について述べた文章（文章ID「1」）や電気自動車の動作について述べた文章（文章ID「2」）などが得られることになる。

【0026】

後述するトピック抽出手段12では、文章を元にトピックを抽出するが、もし、仮にテキストデータを元にトピックを抽出する場合、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抽出した文章を元にトピックを抽出するので、後述するトピック抽出手段12による抽出精度を向上させることができる。

【0027】

次に、共起行列作成手段11は、各文章から第1語群及び第2語群を抽出する。第1語群及び第2語群は、所定の基準により文章から抽出された複数の語からなる。例えば、所定の基準としては、単語や特定の品詞、係り受け表現（文法的構造を持つ単語と単語のペア）などが挙げられる。第1語群と第2語群とで、異なる基準を用いるようにする。このような第1語群及び第2語群は、公知の形態素解析手法あるいは構文解析手法を適用することで得ることができる。

【0028】

次に、共起行列作成手段11は、第1語群に属する語と、第2語群に属する語との組み合わせである共起ペアを含む文章の頻度を計算する。そして、その頻度を要素とする実測共起行列を作成する。実測共起行列のi行j列の要素(i, j)は、第1語群に属するi番目の語と、第2語群に属するj番目の語からなる共起ペアを含む文章の頻度となる。

【0029】

【表2】

		Y:係り受け									
		1,350	539	529	494	444	368	306	289	282	280
		電力-供給	否-判定	バッテリー-充電	モーター-駆動	効率-良い	供給-電力	充電-行う	電気自動車-提供	並列-接続	モーター-供給
X:単語	5,880 構成	118	33	33	36	24	32	25	10	46	30
	5,188 モーター	239	61	58	494	31	54	19	33	27	280
	5,092 制御	268	73	85	108	12	115	41	2	40	74
	4,655 電気自動車	193	73	129	79	59	53	114	289	13	54
	4,604 配置	69	2	8	29	6	15	9	5	12	18
	4,602 バッテリー	337	87	529	60	44	81	72	20	70	106
	4,006 形成	31	4	8	20	1	7	5	0	12	4
	3,978 供給	1,350	43	85	53	10	368	43	7	43	280
	3,960 検出	134	99	36	56	4	44	19	8	27	33
	3,629 電力	1,350	50	127	75	35	368	57	20	36	189

【0030】

10

20

30

50

表 2 に、実測共起行列を例示する。この実測共起行列は、文章から「単語」を抽出して第 1 語群とし、文章から「係り受け表現」を抽出して第 2 語群とするものである。第 1 語群に属する単語として「構成」「モータ」「制御」などが行方向に並び、第 2 語群に属する係り受け表現として「電力-供給」「否-判定」「バッテリー-充電」などが列方向に並んでいる。共起行列作成手段 1 1 は、「構成」と「電力-供給」の共起ペアを含む文章の数をカウントする。表 2 の実測共起行列の例では、要素 (1, 1) の「1 1 8」は、「構成」及び「電力-供給」という共起ペアが存在する文章の頻度 (件数) が 1 1 8 件であることを表している。

【0031】

なお、第 1 語群と第 2 語群の選び方は上述の例に限定されない。例えば、テキストデータ中に含まれる「名詞」に分類される語を第 1 語群とし、「動詞又は形容詞」に分類される語を第 2 語群としてもよい。この第 2 語群のように複数の品詞の何れかに分類される語から第 1 語群又は第 2 語群を抽出してもよい。

【0032】

期待共起行列とは、第 1 語群に属する語及び第 2 語群に属する語の共起ペアの期待頻度を要素とする行列である。期待頻度とは、理論的に推定される共起ペアを含む文章の頻度である。

第 1 語群に属する i 番目の語 (X<sub>i</sub>) が含まれる文章の件数を総頻度 (n (X<sub>i</sub>)) とする。

第 2 語群に属する j 番目の語 (Y<sub>j</sub>) が含まれる文章の件数を総頻度 (n (Y<sub>j</sub>)) とする。

文章の全件数を総文章数 N とする。

期待頻度は、 $n (X_i) \cdot n (Y_j) / N$  である。

【0033】

共起行列作成手段 1 1 は、第 1 語群に属する語が含まれる文章の件数を計上して総頻度 (n (X<sub>i</sub>)) を求め、第 2 語群に属する語が含まれる文章の件数を計上して総頻度 (n (Y<sub>j</sub>)) を求める。そして、文章の全件数を計上して総文章数 N とし、期待頻度を計算する。このような期待頻度を、第 1 語群に属する語及び第 2 語群に属する語からなる全ての共起ペアについて計算する。

【0034】

【表 3】

		Y: 係り受け										
		総頻度 n(Y)	1,350	539	529	494	444	368	306	289	282	280
X: 単語		総頻度 n(X)	電力-供給	否-判定	バッテリー-充電	モータ-駆動	効率-良い	供給-電力	充電-行う	電気自動車-提供	並列-接続	モータ-供給
	5.880	構成	34.6	13.8	13.5	12.7	11.4	9.4	7.8	7.4	7.2	7.2
	5.188	モータ	30.5	12.2	12.0	11.2	10.0	8.3	6.9	6.5	6.4	6.3
	5.092	制御	29.9	12.0	11.7	11.0	9.8	8.2	6.8	6.4	6.3	6.2
	4.655	電気自動車	27.4	10.9	10.7	10.0	9.0	7.5	6.2	5.9	5.7	5.7
	4.604	配置	27.1	10.8	10.6	9.9	8.9	7.4	6.1	5.8	5.7	5.6
	4.602	バッテリー	27.1	10.8	10.6	9.9	8.9	7.4	6.1	5.8	5.7	5.6
	4.006	形成	23.6	9.4	9.2	8.6	7.7	6.4	5.3	5.0	4.9	4.9
	3.978	供給	23.4	9.3	9.2	8.6	7.7	6.4	5.3	5.0	4.9	4.9
	3.960	検出	23.3	9.3	9.1	8.5	7.7	6.3	5.3	5.0	4.9	4.8
	3.629	電力	21.3	8.5	8.4	7.8	7.0	5.8	4.8	4.6	4.5	4.4

【0035】

表 3 は、共起行列作成手段 1 1 により作成された期待共起行列の一例である。総文章数 N は、2 2 9, 5 9 8 件である。第 1 語群の一番目の語 (X 1) である「構成」の総頻度 (n (X 1)) は、5, 8 8 0 件である。第 2 語群の一番目の語 (Y 1) である「電力-供給」の総頻度 (n (Y 1)) は、1, 3 5 0 件である。要素 (1, 1) は「3 4. 6」である。これは、「構成」と「電力-供給」からなる共起ペアを含む文章の頻度は、理論的には「3 4. 6」であることを表している。

【0036】

10

20

30

50

共起行列は、第1語群に属する語、及び第2語群に属する語の組み合わせの頻度に基づく要素からなる行列である。より具体的には、共起行列は、期待共起行列の各要素に対する実測共起行列の各要素の差分あるいは比率を各要素とする行列である。

【0037】

共起行列作成手段11は、期待共起行列の各要素に対する実測共起行列の各要素の差分あるいは比率を計算して共起行列を作成する。この共起行列は、次のトピック抽出手段12の入力データとなる。期待共起行列の各要素に対する実測共起行列の各要素の差分あるいは比率として、実測共起行列の各要素(i, j) / 期待共起行列の各要素(i, j)の対数を計算し、その値を共起行列の要素(i, j)とする。実測共起行列及び期待共起行列の各要素(i, j)がゼロの場合や、上記対数が負であれば、共起行列の要素はゼロとする。このようにして作成した共起行列を表4に例示する。

10

【0038】

なお、差分あるいは比率の取り方は、単純な差分(絶対誤差)としてもよいし、絶対誤差を期待頻度で除した相対誤差としてもよいし、単純な比率としてもよいし、そうした差分あるいは比率の絶対値を取ったり、二乗を取ったり、対数を取ったりしてもよい。ただしゼロで除して値が計算不可となることや値が負数となることがないように、そのような場合は上記のようにゼロに置換するなどの調整を施す。

【0039】

【表4】

		Y: 係り受け										
		総頻度 n(Y)	1,350	539	529	494	444	368	306	289	282	280
X: 単語	総頻度 n(X)		電力-供給	否-判定	バッテリー-充電	モーター-駆動	効率-良い	供給-電力	充電-行う	電気自動車-提供	並列-接続	モーター-供給
	5880	構成	1.2	0.9	0.9	1.0	0.7	1.2	1.2	0.3	1.9	1.4
	5188	モーター	2.1	1.6	1.6	3.8	1.1	1.9	1.0	1.6	1.4	3.8
	5092	制御	2.2	1.8	2.0	2.3	0.2	2.6	1.8	0.0	1.9	2.5
	4655	電気自動車	2.0	1.9	2.5	2.1	1.9	2.0	2.9	3.9	0.8	2.3
	4604	配置	0.9	0.0	0.0	1.1	0.0	0.7	0.4	0.0	0.8	1.2
	4602	バッテリー	2.5	2.1	3.9	1.8	1.6	2.4	2.5	1.2	2.5	2.9
	4006	形成	0.3	0.0	0.0	0.8	0.0	0.1	0.0	0.0	0.9	0.0
	3978	供給	4.1	1.5	2.2	1.8	0.3	4.1	2.1	0.3	2.2	4.1
	3960	検出	1.8	2.4	1.4	1.9	0.0	1.9	1.3	0.5	1.7	1.9
	3629	電力	4.1	1.8	2.7	2.3	1.6	4.1	2.5	1.5	2.1	3.8

【0040】

トピック抽出手段12は、前記共起行列を入力とし、第1語群に属する語及び第2語群に属する語で構成される複数のトピックを抽出する潜在意味解析法を実行することにより、各トピックを条件とした第1語群に属する語の第1条件付確率、及び各トピックを条件とした第2語群に属する語の第2条件付確率を求める。トピックは、発明に関する文章の主題を表しているといえる。

30

【0041】

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。

40

【0042】

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の間に潜在的なトピックZ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トピックZを抽出するものである。PLSAによるトピック抽出は、各トピックZに属する文章Dの条件付確率及び各トピックZに属する単語Wの条件付確率及びトピックZの確率がEMアルゴリズムにより計算される。

50

【0043】

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語群に属する語W1と、第2語群に属する語W2との間に潜在的なトピックZがあると想定し、第1語群に属する語W1と第2語群に属する語W2の組み合わせで構成されるトピックZを抽出するものである。すなわち、トピック抽出手段12は、共起行列を入力としてPLSAを実行することで、各トピックZを条件とした第1語群に属する語W1の第1条件付確率として $P(W1|Z)$ 、及び各トピックZを条件とした第2語群に属する語W2の第2条件付確率として $P(W2|Z)$ を計算する。本実施形態の例では、第1語群に属する語として単語(名詞、動詞、形容詞)を、第2語群に属する語として係り受け表現(名詞と動詞・形容詞の係り受けペア)を設定している。PLSAの具体的な計算方法は、「Hofmann, T.: Probabilistic latent semantic analysis, Proc. Of Uncertainty in Artificial Intelligence, pp.289-296, 1999.」などの文献に記載の公知の技法を用いて実行することができる。

10

【0044】

表5に、PLSAにより計算されたトピックに属する単語及び係り受け表現を例示する。表5には、複数作成されたトピックのうち、2つのトピックZ08とトピックZ21に属する単語及び係り受け表現が示されている。それぞれ条件付確率が高い順に単語および係り受け表現を並べており、それぞれの総頻度 $n(X_i)$ と総頻度 $n(Y_1)$ も掲載している。

20

【0045】

【表5】

トピックZ08					
P(XZ)	n(X)	X: 単語	P(YZ)	n(Y)	Y: 係り受け
2.2%	112	マスタシリンダ	3.2%	32	基づく-発生
2.0%	73	ブレーキ液圧	2.6%	32	操作量-応ずる
1.6%	72	ブレーキ操作	2.6%	33	ブレーキ液圧-発生
1.6%	117	液圧	2.5%	53	制動力-発生
1.4%	779	ブレーキ	2.3%	49	ブレーキ-備える
1.4%	279	制動力	2.1%	20	備える-ブレーキ
1.3%	111	操作量	2.0%	92	車両-ブレーキ
...	...	...	...	...	...

トピックZ21					
P(XZ)	n(X)	X: 単語	P(YZ)	n(Y)	Y: 係り受け
3.2%	180	非接触	3.0%	52	電力-受電
2.5%	78	送電コイル	2.7%	28	給電-電力
2.4%	160	受電	2.3%	35	受電-電力
2.4%	86	受電コイル	2.1%	25	電力-給電
2.2%	329	給電装置	1.8%	20	駐車装置-変化
2.0%	73	受電装置	1.7%	34	非接触-受電
1.8%	124	受電部	1.6%	21	供給-給電装置
...	...	...	...	...	...

【0046】

トピックZ08についてみると、第1条件付確率が最上位である単語は「マスタシリンダ」という単語であり、第2条件付確率が最上位である係り受け表現は「基づく-発生」である。このようなトピックZ08に所属する単語及び係り受け表現に基づいて、トピックZ08の意味を解釈することができる。例えば、トピックZ08は、第1条件付確率が上位である単語に基づけば、ブレーキに関するトピックであると解釈することができる。また各単語および係り受け表現の総頻度にも着目すると、例えば「マスタシリンダ」「ブ

50

レーキ液圧」「ブレーキ操作」「液圧」など、「ブレーキ」という単語よりも比較的頻度の少ないブレーキに関する単語も上位の条件付確率が割り当てられており、より具体的な表現で構成された個性的なトピックが抽出されていることが分かる。

【0047】

PLSAは、トピック数を予め設定する必要があるが、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、本実施形態のトピック抽出手段12では、トピック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トピック数で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、トピック数は、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。

【0048】

本実施形態では、表6に示すように、トピック抽出手段12により50個のトピックが抽出され、それぞれのトピックの解釈がなされた。表6にトピック抽出手段により抽出されたトピックに解釈を与えたものを例示する。

【表6】

No.	トピック名	No.	トピック名
Z01	エンジン制御	Z26	電流・電圧の検出
Z02	動力伝達	Z27	温度、電流、充電量などの検出と制御
Z03	差動機構などを備えた動力伝達の制御	Z28	演算や推定、測定などのステップを含む方法
Z04	回転運動	Z29	情報の取得・提供(位置情報やバッテリー残量等)
Z05	ロータ・ステータなどモータの構成	Z30	スイッチなど操作装置
Z06	モータ制御(トルク制御や回転数制御等)	Z31	車両用灯具
Z07	油圧ポンプなどを利用したモータ駆動	Z32	掃除機
Z08	ブレーキ	Z33	基板の構成
Z09	状態に応じた制御、運転者の操作補助	Z34	回路の接続(電力変換回路等)
Z10	コンバータとバッテリー昇降圧	Z35	端子接続
Z11	直流と交流の電力変換	Z36	部品・装置の収容ケース・筐体
Z12	回転力などの電気エネルギー変換	Z37	部品・装置の配置
Z13	エネルギー効率の向上	Z38	パーツなどの移動、位置
Z14	発電と蓄電	Z39	構造の形成・方位
Z15	電池モジュールの提供	Z40	支持構造
Z16	燃料電池	Z41	装置やユニットの構成
Z17	二次電池の構成	Z42	システム・方法の構成
Z18	バッテリーの充放電	Z43	その他方法
Z19	充電システム	Z44	組成物の製造方法(樹脂や電解液等)
Z20	充電の接続	Z45	機能的組成物・成形品(耐熱性や耐衝撃性等)
Z21	非接触など受給電装置	Z46	製造の効率化(小型化や低コスト化等)
Z22	車両用空調など熱交換	Z47	不具合の防止(損傷、感電、盗難等)
Z23	冷却装置と放熱	Z48	その他
Z24	信号の入出力と検出	Z49	タービン発電と船舶・飛行機への応用
Z25	電気信号の取得と変換(センサ検出等)	Z50	重力発電の活用による地球温暖化防止

【0049】

スコア計算手段13は、第1条件付確率及び第1語群の出現頻度、並びに第2条件付確率及び第2語群の出現頻度に基づいて、各トピックを条件とした各文章の条件付確率を計算する。そして、この条件付確率を各文章の発生確率で除した値を、各文章に対する各トピックのスコアとする。そして、その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トピックのスコアを求める。

【0050】

文章  $S_h$  におけるトピック  $Z_k$  のスコアは、 $P(S_h | Z_k) / P(S_h)$  である (式 (1))。  $k$  は、PLSA で作成されたトピックを特定する番号であり、トピック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。

【0051】

【数1】

$$(1) \frac{P(S_h | Z_k)}{P(S_h)}$$

【0052】

第1語群に含まれる語 (行要素  $X_i$ ) の集合を  $S_{x_h}$  とし (式 (2))、第2語群に含まれる語 (列要素  $Y_j$ ) の集合を  $S_{y_h}$  とする (式 (3))。

【0053】

【数2】

$$(2) S_{x_h} = \{X_1, X_2, \dots, X_i\}$$

$$(3) S_{y_h} = \{Y_1, Y_2, \dots, Y_j\}$$

【0054】

式 (1) の  $P(S_h | Z_k)$  は、上記文章  $S_{x_h}$  と文章  $S_{y_h}$  に分解し、それぞれ  $P(S_{x_h} | Z_k)$  と  $P(S_{y_h} | Z_k)$  を計算し、それらを統合して  $P(S_h | Z_k)$  を計算する。

【0055】

トピック  $Z_k$  を条件とした文章  $S_{x_h}$  の条件付確率  $P(S_{x_h} | Z_k)$  を計算し (式 (4))、トピック  $Z_k$  を条件とした文章  $S_{y_h}$  の条件付確率  $P(S_{y_h} | Z_k)$  を計算する (式 (5))。

【0056】

【数3】

$$(4) P(S_{x_h} | Z_k) = \sum_i P(S_{x_h} | X_i) P(X_i | Z_k)$$

$$(5) P(S_{y_h} | Z_k) = \sum_j P(S_{y_h} | Y_j) P(Y_j | Z_k)$$

【0057】

式 (4) の行要素  $X_i$  が出現する中で文章  $S_{x_h}$  が出現する確率 (第1語群の出現頻度) である  $P(S_{x_h} | X_i)$  は、 $X_i$  が出現する総頻度  $n(X_i)$  の逆数として計算される (式 (6))。

【0058】

【数4】

$$(6) P(S_{x_h} | X_i) = 1/n(X_i)$$

【0059】

式 (5) の列要素  $Y_j$  が出現する中で文章  $S_{y_h}$  が出現する確率 (第2語群の出現頻度) である  $P(S_{y_h} | Y_j)$  は、 $Y_j$  が出現する総頻度  $n(Y_j)$  の逆数として計算される (式 (7))。

【0060】

【数5】

$$(7) P(Sy_h|Y_j) = 1/n(Y_j)$$

【0061】

式(4)、式(5)のトピック $Z_k$ を条件とした行要素 $X_i$ の条件付確率(第1条件付確率)である $P(X_i|Z_k)$ と、トピック $Z_k$ を条件とした列要素 $Y_j$ の条件付確率(第2条件付確率)である $P(Y_j|Z_k)$ は、PLSAの実行で得られる。したがって、式(1)のトピック $Z_k$ を条件とした文章 $S_h$ の条件付確率 $P(S_h|Z_k)$ は、式(8)

10

【0062】

【数6】

$$(8) P(S_h|Z_k) = P(S_h|Sx_h)P(Sx_h|Z_k) + P(S_h|Sy_h)P(Sy_h|Z_k)$$

【0063】

文章 $S_h$ において、行要素 $X$ で定義される文章 $S_{x_h}$ と、列要素 $Y$ で定義される文章 $S_{y_h}$ の重みは同じであるため、式(8)中の、文章 $S_{x_h}$ を条件とした文章 $S_h$ の条件付確率 $P(S_h|S_{x_h})$ と、文章 $S_{y_h}$ を条件とした文章 $S_h$ の条件付確率 $P(S_h|S_{y_h})$ はそれぞれ0.5とする。

20

【0064】

式(1)の文章 $S_h$ の確率 $P(S_h)$ は、式(9)で表され、 $P(Z_k)$ はPLSAの実行で得られる。

【0065】

【数7】

$$(9) P(S_h) = \sum_k P(S_h|Z_k)P(Z_k)$$

【0066】

このように、 $P(S_h|Z_k)$ と $P(S_h)$ との比をもって文章 $S_h$ におけるトピック $Z_k$ のスコアとする。この値が1を超えるということは、文章 $S_h$ の発生確率はトピック $Z_k$ を条件とすることで上昇し、トピック $Z_k$ との関係が強いということである。このようなスコアを採用することで、各文章 $S_h$ とトピック $Z_k$ の関係の強さを把握しやすくなることができる。表7に各文章 $S_h$ に対する各トピック $Z_k$ のスコアを例示する。

【0067】

【表7】

		トピックZ			
テキストデータID	文章ID(h)	Z1	Z2	...	Z50
1	1	$P(S1 T1)/P(S1)=3.1$	$P(S1 T2)/P(S1)=0.9$		$P(S1 T14)/P(S1)=1.1$
1	2	$P(S2 T1)/P(S2)=1.4$	$P(S2 T2)/P(S2)=0.2$		$P(S2 T14)/P(S2)=2.4$
1	3	$P(S3 T1)/P(S3)=0.8$	$P(S3 T2)/P(S3)=5.8$		$P(S3 T14)/P(S3)=0.9$
1	4	$P(S4 T1)/P(S4)=1.2$	$P(S4 T2)/P(S4)=3.2$		$P(S4 T14)/P(S4)=1.0$
2	5	$P(S5 T1)/P(S5)=0.6$	$P(S5 T2)/P(S5)=1.8$		$P(S5 T14)/P(S5)=1.6$

【0068】

例えば、文章ID「1」は、トピックZ1についてのスコアが3.1であり、トピックZ2についてのスコアが0.9であり、このようなスコアが全トピックについて計算され

50

ている。

【0069】

スコア計算手段13は、文章ID単位に計算された各トピックのスコアをテキストデータID単位に集約する。文章単位のスコアをテキストデータ単位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。本実施形態では、トピック毎のスコアの最大値を、テキストデータIDの各トピックのスコアとする。

【0070】

【表8】

テキストデータID	文章ID(h)	トピックZk			
		Z1	Z2	...	Z50
1	1	<i>3.1</i>	<i>0.9</i>		<i>1.1</i>
1	2	<i>1.4</i>	<i>0.2</i>		<i>2.4</i>
1	3	<i>0.8</i>	<i>5.8</i>		<i>0.9</i>
1	4	<i>1.2</i>	<i>3.2</i>		<i>1.0</i>

【0071】

表8を用いて、スコアの集計について具体的に説明する。テキストデータ「1」は文章「1」～文章「4」から構成されている。トピックごとに、文章「1」～文章「4」のうち最大値を求める。

【0072】

文章「1」～文章「4」に対するトピックZ1のスコアは「3.1」「1.4」「0.8」「1.2」である。したがって、「3.1」が最大値となる。この最大値「3.1」がテキストデータ「1」に対するトピックZ1のスコアとなる。以下同様に、トピックZ2～Z50についてトピック毎に最大値を計算することで、テキストデータ「1」に対する各トピックのスコアを得る。このような最大値を求めてテキストデータに対する各トピック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トピックのスコアである。このようにして、各テキストデータに対して、各トピックのスコアを得ることができる。

【0073】

このようにして得られたスコアから、トピックの該当の有無を表す1, 0の情報を付与してもよい。例えば、閾値を「3」に設定し、スコアが3以上であれば「1」に3未満であれば「0」というフラグ情報を付与してもよい。表9にフラグ情報を示す。

【0074】

【表9】

テキストデータID	トピックZk			
	Z1	Z2	...	Z50
1	1	1		0
2	0	1		0

【0075】

テキストデータ「1」は、トピックZ1のスコアが「3.1」であるから（表9参照）、フラグ情報は「1」となる。同様に、トピックZ2のスコアは「5.8」であるから、フラグ情報は「1」となる。トピックZ50のスコアは「2.4」であるから、フラグ情報は「0」となる。なお、閾値は「3」である必要はない。 $P(S_h | Z_k) / P(S_h)$ で定義したスコアは1が基準と考えることができるので、閾値を「1」と設定してもよい。

【0076】

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3は、分析装置での処理を示すフローチャートである。

**【0077】**

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：共起行列作成ステップ）。具体的には、共起行列作成手段11が、テキストデータから文章を抽出し、各文章に含まれている第1語群に属する語及び第2語群に属する語の組み合わせの個数を表す共起行列を作成し、これは実測共起行列と期待共起行列とから作成する。具体例については、上述したので説明は省略する。

**【0078】**

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：トピック抽出ステップ）。具体的には、トピック抽出手段12が共起行列を入力とし、第1語群に属する語及び第2語群に属する語で構成される複数のトピックを抽出する潜在意味解析法を実行する。これにより、各トピックを条件とした第1語群に属する語の第1条件付確率、及び各トピックを条件とした第2語群に属する語の第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。

10

**【0079】**

次に、各テキストデータに対する各トピック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語群の出現頻度、並びに第2条件付確率及び第2語群の出現頻度に基づいて、各トピックを条件とした各文章の条件付確率を各文章に対する各トピック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トピックのスコアを求める。具体例については上述したので説明は省略する。

20

**【0080】**

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トピックを抽出し、各テキストデータに対してトピックのスコアを求める。このようなスコアを求める前提となる共起行列は、期待共起行列に対する実測共起行列の差分あるいは比率を元に得られている。

**【0081】**

このようにして得られた共起行列を用いることで、テキストデータからより個性的なトピックを抽出することができる。これは次のような理由による。実測共起行列の各要素を実測共起頻度、期待共起行列の各要素を期待共起頻度と称する。実測共起頻度が高い共起ペアでも、元々全体の頻度が高い要素（表2でいう総頻度が高い第1語群の語や第2語群の語）が含まれるときには期待共起頻度も高くなるため、実測共起頻度を期待共起頻度で除すことで期待頻度の大きさが制限される。逆に実測共起頻度が高くない共起ペアでも、期待共起頻度がそれよりも十分低ければ共起行列の期待頻度は大きくなり、これにPLSAを適用した解ではこうした要素にも高い確率が割り当てられる可能性がある。つまり、通常のPLSAでは頻度が低い要素は高い確率が割り当てられない傾向があるが、上述したような共起行列を用いる本発明では、そうした要素にも高い確率が割り当てられる可能性があり、より個性的なトピックが抽出されることが期待できる。

30

**【0082】**

なお、通常の共起行列を用いてPLSAを適用した場合、頻度が高い要素に高い確率が割り当てられることから、結果として抽出されるトピックは典型的なものになる傾向があり、目新しさに欠けてしまう。

40

**【0083】**

また、テキストデータに含まれる文章ごとに共起行列を作成し、トピック抽出手段12により文章を元にトピックを抽出した。これにより、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場合であっても、トピック抽出手段12による抽出されたトピックは、異なる観点が混在したような曖昧さが低減され、より明確な内容のトピックを抽出することができる。

**【0084】**

また、共起行列の各要素は、期待共起行列に対する実測共起行列の比率の対数とした。このように対数を用いることにより、共起行列の比率が極端に高くなることを制限するこ

50

とができる。特に期待共起頻度は1未満となるケースも多く、比率のみでは値が高くなりすぎるものもある。この状態では共起行列全体の値の分布は大きくばらつき、極端な値の開きが生まれてしまうため、PLSAを適用した際の最適化計算において、今度はこの極端に大きな値に引っ張られる結果となり、必要以上にデフォルメされた歪んだトピックとなることがありうる。そこで、この比率の値の対数を取ることで値の分布をならし、上記の現象を制限し、より適正なトピックを得ることができると期待できる。なお、共起行列の各要素の値は、期待共起行列に対する実測共起行列の差分あるいは比率を取ることで計算されるが、この差分あるいは比率の取り方は、単純な差分（絶対誤差）としてもよいし、絶対誤差を期待頻度で除した相対誤差としてもよいし、単純な比率としてもよいし、そうした差分あるいは比率の絶対値を取ったり、二乗を取ったり、対数を取ったりしてもよい。

10

**【0085】**

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、一台の分析装置1において各手段11～13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。

**【0086】**

また、上記実施形態では、特許文献を対象としたものであるが、これに限定されない。例えば、顧客から得たアンケートの自由記述結果をテキストデータとし、顧客の潜在ニーズを抽出したり、コールセンターの問い合わせ履歴をテキストデータとし、消費者の隠れた評価の観点を抽出するなど、テキストデータの一般に適用することができる。

20

**【0087】**

〈比較例〉

上述した実施形態と同じテキストデータを用いて、実測共起行列及び期待共起行列を作成せずに、実測共起行列を共起行列としてトピックの抽出及びスコアの集計を行った比較例を示す。具体的には、テキストデータから文章を抽出し、各文章から、第1語群及び第2語群を抽出し、各文章に含まれている第1語群に属する語及び第2語群に属する語の組み合わせの個数を表す共起行列を作成する。

**【0088】**

このようにして作成した共起行列について、上述した実施形態と同様にトピック抽出を行った結果を表10に示す。本発明では表6に示したように、50個のトピックが抽出されたが、比較例においては表10に示すように、34個のトピックが抽出された。

30

**【0089】**

【表 10】

No.	トピック名	No.	トピック名
Z01	エンジンの始動と停止	Z18	演算・推定
Z02	動力の伝達	Z19	機器の異常検出
Z03	モータ駆動	Z20	操作スイッチ
Z04	ロータ・ステータなど回転部品の構成	Z21	筐体
Z05	ブレーキ装置	Z22	表面の形成
Z06	動作制御	Z23	位置とその移動
Z07	動力伝達の制御	Z24	配置・位置・方向
Z08	スイッチの切り替え	Z25	構成の方位
Z09	交流・直流の変換	Z26	構成
Z10	エネルギーの変換	Z27	接続
Z11	電池モジュールの提供	Z28	方法の提供
Z12	二次電池の構成	Z29	損傷や浸水など不具合の防止
Z13	電気自動車の蓄電池充電	Z30	小型化・簡素化・低コスト化など付加価値
Z14	非接触受電など給電装置	Z31	効率性・安全性の向上
Z15	外部への電力供給	Z32	既存エンジンへの警鐘・樹脂組成物の提供
Z16	空調などの冷却・加熱	Z33	重力発電の活用による地球温暖化防止
Z17	情報通信	Z34	タービン発電の出力向上・燃費低減

【0090】

【表 11】

トピックZ09					
P(XZ)	n(X)	X:単語	P(YZ)	n(Y)	Y:係り受け
1.4%	50	シフトレンジ	1.9%	23	自動的-行う
1.0%	38	パーキングレンジ	1.7%	38	操作-行う
0.9%	147	検出結果	1.6%	22	駆動-停止
0.7%	1,200	停止	1.6%	26	動作-行う
0.7%	365	解除	1.6%	40	要する-時間
0.6%	34	キースイッチ	1.4%	46	停止-状態
0.6%	3,960	検出	1.2%	26	ブレーキ-作動
...	...	...	...	...	...

トピックZ29					
P(XZ)	n(X)	X:単語	P(YZ)	n(Y)	Y:係り受け
1.2%	122	ナビゲーション装置	2.1%	48	情報-送信
1.1%	809	情報	2.1%	42	情報-含む
1.0%	165	目的地	1.8%	68	情報-取得
1.0%	111	位置情報	1.5%	31	情報-受信
0.9%	786	取得	1.5%	35	示す-情報
0.9%	819	送信	1.5%	126	情報-基づく
0.8%	558	表示	1.4%	25	情報-用いる
...	...	...	...	...	...

【0091】

本発明で抽出された50個のトピックには、上記比較例で抽出された34個のトピックに対応するものもあるが、上記比較例では抽出されずに、本発明によってのみ得られたトピックも存在した。表11にその例を示す。トピックZ09は、「シフトレンジ」や「パーキングレンジ」、「検出」、「停止」、「自動的-行う」といった表現で確率が高く、運転者の誤操作を抑制したり自動停止などの運転アシストに関する技術と解釈できる。トピックZ29は、「ナビゲーション装置」や「情報」、「目的地」、「位置情報」といった表現で確率が高く、位置情報を取得してドライバーにナビ情報として提供するなど、情

報の取得と提供に関する技術と解釈できる。どちらも近年の自動車業界において付加価値を高める重要な機能が、本発明によってテキストデータから得ることができた。

【0092】

〈実施形態2〉

実施形態1では、複数あるテキストデータのそれぞれから文章を抽出し、各文章から共起行列を作成した。しかしながら、本発明はこれに限定されず、複数あるテキストデータから共起行列を作成してもよい。以下、本実施形態の分析方法、分析装置、分析プログラムについて説明するが、実施形態1と重複する説明は省略する。

【0093】

共起行列作成手段11は、テキストデータから第1語群に属する語及び第2語群に属する語の組み合わせの頻度を表す共起行列を作成する。つまり、テキストデータは1又は複数の文章からなるが、文章単位では処理せずに、テキストデータ単位で処理する。なお、例として用いるテキストデータは、実施形態1の表1と同様である。

10

【0094】

まず、共起行列作成手段11は、各テキストデータから第1語群及び第2語群を抽出する。

【0095】

次に、共起行列作成手段11は、第1語群に属する語と、第2語群に属する語との組み合わせである共起ペアを含むテキストデータの頻度を計算する。そして、その頻度を要素とする実測共起行列を作成する。実測共起行列の*i*行*j*列の要素(*i*, *j*)は、第1語群に属する*i*番目の語と、第2語群に属する*j*番目の語からなる共起ペアを含むテキストデータの頻度となる。

20

【0096】

次に、共起行列作成手段11は、第1語群に属する語が含まれるテキストデータの件数を計上して総頻度( $n(X_i)$ )を求め、第2語群に属する語が含まれるテキストデータの件数を計上して総頻度( $n(Y_j)$ )を求める。そして、テキストデータの全件数を計上して総テキストデータ数*N*とし、期待頻度を計算する。このような期待頻度を、全ての第1語群に属する語及び第2語群に属する語について計算し、期待共起行列を作成する。

【0097】

次に、共起行列作成手段11は、期待共起行列の各要素に対する実測共起行列の各要素の差分あるいは比率を計算して共起行列を作成する。実施形態1と同様に実測共起行列の各要素(*i*, *j*) / 期待共起行列の各要素(*i*, *j*)の対数を計算し、その値を共起行列の要素(*i*, *j*)とする。

30

【0098】

このようにして得られた共起行列に対して、トピック抽出手段12によりトピックの抽出を行う。この抽出については、実施形態1と同様であるのでここでの説明は省略する。

【0099】

実施形態1では、各トピックを条件とした各文章の条件付確率を計算したが、本実施形態では、各トピックを条件とした各テキストデータの条件付確率を計算する。

【0100】

具体的には、スコア計算手段13は、第1条件付確率及び第1語群の出現頻度、並びに第2条件付確率及び第2語群の出現頻度に基づいて、各トピックを条件とした各テキストデータの条件付確率を計算する。そして、この条件付確率を各テキストデータの発生確率で除した値を、各テキストデータに対する各トピックのスコアとする。

40

【0101】

テキストデータ $D_h$ におけるトピック $Z_k$ のスコアは、 $P(D_h | Z_k) / P(D_h)$ である(式(10))。kは、PLSAで作成されたトピックを特定する番号であり、トピックの総数を最大とする自然数である。hは、テキストデータを特定する番号(テキストデータID)であり、テキストデータの総数を最大とする自然数である。

【0102】

50

【数 8】

$$(10) \frac{P(D_h|Z_k)}{P(D_h)}$$

【0103】

第1語群に含まれる語（行要素 $X_i$ ）の集合を $Dx_h$ とし（式（11））、第2語群に含まれる語（列要素 $Y_j$ ）の集合を $Dy_h$ とする（式（12））。

【0104】

【数 9】

$$(11) Dx_h = \{X_1, X_2, \dots, X_i\}$$

$$(12) Dy_h = \{Y_1, Y_2, \dots, Y_j\}$$

【0105】

これらの集合を用いて、トピック $Z_k$ を条件としたテキストデータ $Dx_h$ の条件付確率 $P(Dx_h|Z_k)$ を計算し（式（13））、トピック $Z_k$ を条件としたテキストデータ $Dy_h$ の条件付確率 $P(Dy_h|Z_k)$ を計算する（式（14））。

【0106】

【数 10】

$$(13) P(Dx_h|Z_k) = \sum_i P(Dx_h|X_i)P(X_i|Z_k)$$

$$(14) P(Dy_h|Z_k) = \sum_j P(Dy_h|Y_j)P(Y_j|Z_k)$$

【0107】

式（13）の行要素 $X_i$ が出現する中でテキストデータ $Dx_h$ が出現する確率（第1語群の出現頻度）である $P(Dx_h|X_i)$ は、 $X_i$ が出現する総頻度 $n(X_i)$ の逆数として計算される（式（15））

【0108】

【数 11】

$$(15) P(Dx_h|X_i) = 1/n(X_i)$$

【0109】

式（14）の列要素 $Y_j$ が出現する中でテキストデータ $Dy_h$ が出現する確率（第1語群の出現頻度）である $P(Dy_h|Y_j)$ は、 $Y_j$ が出現する総頻度 $n(Y_j)$ の逆数として計算される（式（16））

【0110】

【数 12】

$$(16) P(Dy_h|Y_j) = 1/n(Y_j)$$

【0111】

式（13）、式（14）のトピック $Z_k$ を条件とした行要素 $X_i$ の条件付確率（第1条件付確率）である $P(X_i|Z_k)$ と、トピック $Z_k$ を条件とした列要素 $Y_j$ の条件付確率（第2条件付確率）である $P(Y_j|Z_k)$ は、PLSAの実行で得られる。したがって、式（10）のトピック $Z_k$ を条件としたテキストデータ $D_h$ の条件付確率 $P(D_h|Z_k)$ は、式（17）で表される。

10

20

30

40

50

【0112】

【数13】

$$(17) P(D_h|Z_k) = P(D|Dx_h)P(Dx_h|Z_k) + P(D_h|Dy_h)P(Dy_h|Z_k)$$

【0113】

テキストデータ  $D_h$  において、行要素  $X$  で定義される文章  $D_{hx}$  と、列要素  $Y$  で定義されるテキストデータ  $D_{yh}$  の重みは同じであるため、式 (17) 中の、テキストデータ  $D_{xh}$  を条件としたテキストデータ  $D_h$  の条件付確率  $P(D_h|D_{xh})$  と、テキストデータ  $D_{yh}$  を条件としたテキストデータ  $D_h$  の条件付確率  $P(D_h|D_{yh})$  はそれぞれ 0.5 とする。

10

【0114】

式 (10) のテキストデータ  $D_h$  の確率  $P(D_h)$  は、式 (18) で表され、 $P(Z_k)$  は PLSA の実行で得られる。

【0115】

【数14】

$$(18) P(D_h) = \sum_k P(D_h|Z_k)P(Z_k)$$

【0116】

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、実施形態 1 と同様の作用効果を奏する。また、本実施形態では、文章ごとではなく、テキストデータから共起行列を作成する。このため、本実施形態の分析方法等は、テキストデータに異なる観点の文章が複数含まれていない場合に、特に有用である。

【0117】

〈実施形態 3〉

実施形態 1 ではテキストデータから抽出された文章を対象として共起行列を作成し、実施形態 2 ではテキストデータを対象として共起行列を作成したが、本発明はこれらに限定されない。

30

【0118】

本実施形態のテキストデータは、カテゴリに分類されたテキスト部 (1 又は複数の文章からなる) を複数備えた構造となっている。表 12 にテキストデータを例示する。

【0119】

【表 1 2】

テキストデータID	カテゴリ	テキスト部
1	タイトル	電気自動車
	課題	電気自動車の充電完了後における無駄な電力消費を抑制する。
	解決手段	制御部は、計測部の計測結果から電気自動車の充電完了を判断し、開閉部を制御して給電路を開成させる。
	効果	その結果、電気自動車の充電完了後における無駄な電力消費を抑制することができる。
2	タイトル	非接触充電システム
	課題	非接触充電の送電効率の向上を図る
	解決手段	受電コイルを有する受電ユニットの電気自動車への取付構造において、受電ユニットを電気自動車が有する一対の後輪の間に配置する。
	効果	非接触充電の送電効率が向上する。

## 【0120】

表 1 2 に示すように、テキストデータは、複数のテキスト部からなり、各テキスト部は、カテゴリに分類されている。例えば、特許出願の明細書等に関するテキストデータには、タイトル（発明の名称）、課題、解決手段、効果などのカテゴリに分類されたテキスト部が含まれている。

20

## 【0121】

共起行列作成手段 1 1 は、複数のカテゴリのうち特定の 2 個のカテゴリを用いる。この 2 個のカテゴリは、ユーザーに指定されたものである。それらの 2 個のカテゴリのうちの一つを第 1 のカテゴリ、他の一つを第 2 のカテゴリと称する。

## 【0122】

共起行列作成手段 1 1 は、第 1 のカテゴリに分類されたテキスト部から第 1 語群に属する語、及び第 2 のカテゴリに分類されたテキスト部から第 2 語群に属する語の組み合わせの頻度を表す共起行列を作成する。

30

## 【0123】

共起行列作成手段 1 1 は、全てのテキストデータのうち、第 1 のカテゴリに分類されたテキスト部から第 1 語群を抽出し、第 2 のカテゴリに分類されたテキスト部から第 2 語群を抽出する。

## 【0124】

次に、共起行列作成手段 1 1 は、第 1 語群に属する語と、第 2 語群に属する語との組み合わせである共起ペアを含むテキストデータの頻度を計算する。そして、その頻度を要素とする実測共起行列を作成する。実測共起行列の  $i$  行  $j$  列の要素  $(i, j)$  は、第 1 語群に属する  $i$  番目の語と、第 2 語群に属する  $j$  番目の語からなる共起ペアを含むテキストデータの頻度となる。

40

## 【0125】

【表 1 3】

		解決手段									
		基づく-発生	操作量- 応ずる	ブレーキ 液圧-発生	制動力- 発生	ブレーキ- 備える	電力-受 電	非接触- 受電	給電-電 力	駐車装置 -変化	制御部- 備える
タイトル	ブレーキ	8	97	19	9	41	15	56	67	62	80
	制御	28	39	53	54	98	125	59	70	11	75
	充電	198	115	41	57	64	89	89	91	25	102
	非接触	96	62	77	43	19	74	4	140	51	40
	制動	10	14	147	136	57	130	108	18	94	50
	車両	21	1	23	54	150	34	26	63	70	146
	発電	7	13	9	197	179	195	44	61	8	139
	放電	34	83	105	137	178	49	60	83	4	57
	回生	81	46	29	96	58	141	167	181	132	83
バッテリー	156	24	23	196	4	94	37	153	13	57	

## 【0126】

表 1 3 は、第 1 のカテゴリを「タイトル」とし、第 2 のカテゴリを「解決手段」とし、第 1 語群を「名詞」とし、第 2 語群を「係り受け表現」として作成した実測共起行列を例示している。

## 【0127】

例えば、要素 (1, 1) は、第 1 のカテゴリ「タイトル」に分類されたテキスト部に「ブレーキ」という名詞が含まれ、かつ、第 2 のカテゴリ「解決手段」に分類されたテキスト部に「基づく-発生」という係り受け表現が含まれるような共起ペアが存在するテキストデータの数は 8 件であることを表す。

## 【0128】

次に、共起行列作成手段 1 1 は、第 1 のカテゴリに分類されたテキスト部に、第 1 語群に属する語が含まれるテキストデータの件数を計上して総頻度 ( $n(X_i)$ ) を求める。また、共起行列作成手段 1 1 は、第 2 のカテゴリに分類されたテキスト部に、第 2 語群に属する語が含まれるテキストデータの件数を計上して総頻度 ( $n(Y_j)$ ) を求める。そして、テキストデータの全件数を計上して総テキストデータ数  $N$  とし、期待頻度を計算する。このような期待頻度を、全ての第 1 語群に属する語及び第 2 語群に属する語について計算し、期待共起行列を作成する。

## 【0129】

次に、共起行列作成手段 1 1 は、期待共起行列の各要素に対する実測共起行列の各要素の差分あるいは比率を計算して共起行列を作成する。実施形態 1 と同様に実測共起行列の各要素 ( $i, j$ ) / 期待共起行列の各要素 ( $i, j$ ) の対数を計算し、その値を共起行列の要素 ( $i, j$ ) とする。

## 【0130】

このようにして得られた共起行列に対して、トピック抽出手段 1 2 によりトピックの抽出を行う。この抽出については、実施形態 1 と同様であるのでここでの説明は省略する。

## 【0131】

本実施形態におけるスコアの計算は、 $Dx_h$ 、 $Dy_h$  の定義が異なる以外は、実施形態 2 と同様であるので詳細な説明は省略する。 $Dx_h$  は、第 1 のカテゴリに分類されたテキスト部から得られた、第 1 語群に含まれる語 (行要素  $X_i$ ) の集合である (式 (19))。  $Dy_h$  は、第 2 のカテゴリに分類されたテキスト部から得られた、第 2 語群に含まれる語 (列要素  $Y_j$ ) の集合である (式 (20))。

## 【0132】

## 【数 1 5】

$$(19) Dx_h = \{X_1, X_2, \dots, X_i\}$$

$$(20) Dy_h = \{Y_1, Y_2, \dots, Y_j\}$$

## 【0133】

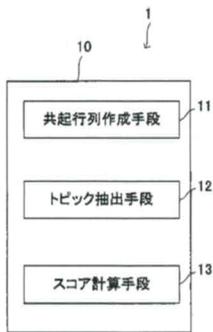
以上に述べたように、本実施形態に係る分析方法、分析装置及び分析プログラムによれば、実施形態1及び実施形態2と同様の作用効果を奏する。また、本実施形態では、カテゴリに分けられたテキスト部を含む、構造化されたテキストデータを対象として分析する場合に特に有用である。

【符号の説明】

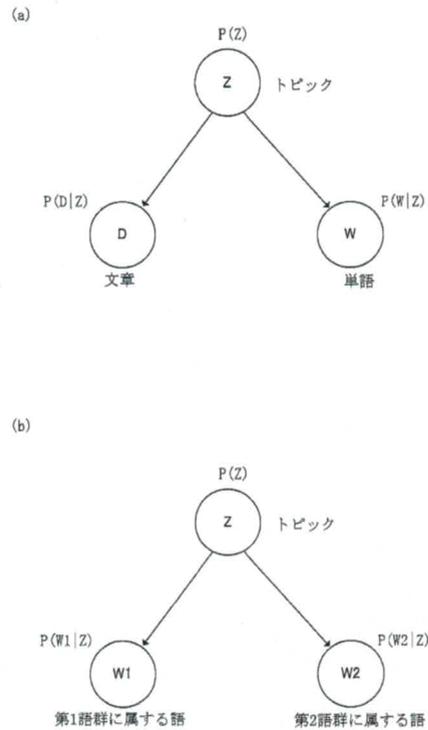
【0134】

- 1 分析装置
- 10 分析プログラム
- 11 共起行列作成手段
- 12 トピック抽出手段
- 13 スコア計算手段

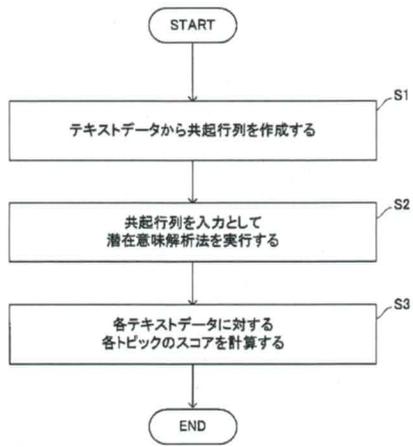
【図1】



【図2】



【図 3】



---

フロントページの続き

(56)参考文献 特開2009-093647 (JP, A)  
特開2002-041543 (JP, A)  
特開2004-288168 (JP, A)

(58)調査した分野(Int.Cl., DB名)

G06F 16/00  
G06F 40/00