

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7221526号
(P7221526)

(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.C1.	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 10 (全 28 頁)

(21)出願番号	特願2019-84331(P2019-84331)	(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
(22)出願日	平成31年4月25日(2019.4.25)	(74)代理人	100101236 弁理士 栗原 浩之
(65)公開番号	特開2019-200784(P2019-200784A)	(74)代理人	100166914 弁理士 山▲崎▼ 雄一郎
(43)公開日	令和1年11月21日(2019.11.21)	(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
審査請求日	令和4年1月20日(2022.1.20)		
(31)優先権主張番号	特願2018-90885(P2018-90885)		
(32)優先日	平成30年5月9日(2018.5.9)		
(33)優先権主張国・地域又は機関	日本国(JP)		
		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及び前記テキストデータに関するメタ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個数を表す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、

前記メタデータが予め設定した事象に該当するとき、当該メタデータに関する前記テキストデータについて第1の共起行列を作成し、

前記メタデータが前記事象に該当しないとき、当該メタデータに関する前記テキストデータ、あるいは前記事象の該当有無にかかわらず全ての前記テキストデータについて第2の共起行列を作成し、

前記第1の共起行列と前記第2の共起行列との差を計算することで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、

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、

前記共起行列作成ステップは、

前記メタデータが予め設定した事象に該当するとき、当該メタデータに関する前記第1のカテゴリに分類された前記テキスト部、及び当該メタデータに関する前記第2のカテゴリに分類された前記テキスト部について第1の共起行列を作成し、

前記メタデータが前記事象に該当しないとき、当該メタデータに関する前記第1のカテゴリに分類された前記テキスト部、及び当該メタデータに関する前記第2のカテゴリに分類された前記テキスト部、あるいは前記事象の該当有無にかかわらず全ての前記テキスト部について第2の共起行列を作成し、

前記第1の共起行列と前記第2の共起行列との差を計算することで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3】

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、

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、

前記共起行列作成ステップは、

前記メタデータが予め設定した事象に該当するとき、当該メタデータに関する前記文章について第1の共起行列を作成し、

前記メタデータが前記事象に該当しないとき、当該メタデータに関する前記文章、あるいは前記事象の該当有無にかかわらず全ての前記文章について第2の共起行列を作成し、

前記第1の共起行列と前記第2の共起行列との差を計算することで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】

請求項1又は請求項2に記載する分析方法において、

前記共起行列作成ステップでは、前記第1の共起行列に対して補正值を乗じ、

前記補正值は、前記第1の共起行列の作成に用いられた前記テキストデータの件数に対する、前記第2の共起行列の作成に用いられた前記テキストデータの件数の比率、あるいは全ての前記テキストデータの件数の比率である

ことを特徴とする分析方法。

【請求項5】

請求項1又は請求項2に記載する分析方法において、

10

20

30

40

50

前記共起行列作成ステップでは、前記第2の共起行列に対して補正值を乗じ、

前記補正值は、前記第2の共起行列の作成に用いられた前記テキストデータの件数に対する、前記第1の共起行列の作成に用いられた前記テキストデータの件数の比率、あるいは全ての前記テキストデータの件数の比率である

ことを特徴とする分析方法。

【請求項6】

請求項3に記載する分析方法において、

前記共起行列作成ステップでは、前記第1の共起行列に対して補正值を乗じ、

前記補正值は、前記第1の共起行列の作成に用いられた前記テキストデータの文章数に対する、前記第2の共起行列の作成に用いられた前記テキストデータの文章数の比率、あるいは全ての前記テキストデータの文章数の比率である

10

ことを特徴とする分析方法。

【請求項7】

請求項3に記載する分析方法において、

前記共起行列作成ステップでは、前記第2の共起行列に対して補正值を乗じ、

前記補正值は、前記第2の共起行列の作成に用いられた前記テキストデータの文章数に対する、前記第1の共起行列の作成に用いられた前記テキストデータの文章数の比率、あるいは全ての前記テキストデータの文章数の比率である

ことを特徴とする分析方法。

【請求項8】

請求項1から請求項7の何れか一項に記載する分析方法において、

前記トピックごとに、

前記スコアが所定の閾値以上である条件の下で前記メタデータが前記事象に該当する確率を前記メタデータが前記事象に該当する確率で除した事象該当ありの指標値、及び

前記スコアが所定の閾値以上である条件の下で前記メタデータが前記事象に該当しない確率を前記メタデータが前記事象に該当しない確率で除した事象該当なしの指標値を計算する集計ステップを備える

ことを特徴とする分析方法。

【請求項9】

テキストデータ、及び前記テキストデータに関するメタデータの分析装置であって、

20

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個数を表す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、

40

前記共起行列作成手段は、

前記メタデータが予め設定した事象に該当するとき、当該メタデータに関する前記テキストデータについて第1の共起行列を作成し、

前記メタデータが前記事象に該当しないとき、当該メタデータに関する前記テキストデータ、あるいは前記事象の該当有無にかかわらず全ての前記テキストデータについて第2の共起行列を作成し、

前記第1の共起行列と前記第2の共起行列との差を計算することで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10】

50

テキストデータ、及び当該テキストデータに関するメタ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個数を表す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、

前記メタデータが予め設定した事象に該当するとき、当該メタデータに関する前記テキストデータについて第1の共起行列を作成し、

前記メタデータが前記事象に該当しないとき、当該メタデータに関する前記テキストデータ、あるいは前記事象の該当有無にかかわらず全ての前記テキストデータについて第2の共起行列を作成し、

前記第1の共起行列と前記第2の共起行列との差を計算することで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。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、事象に影響を与えるトピックをテキストデータから抽出する分析方法、分析装置及び分析プログラムに関する。

【背景技術】

【0002】

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。例えば、顧客満足の要因を探るために、アンケートの自由記述や口コミに代表されるコメントのテキストデータから、ターゲットとなる事象、例えば満足度の評価得点に影響を与える評価内容（トピック）を抽出すること、あるいはサービスの解約や会員退会を防止する要因を探るために、コールセンターなどの問い合わせ履歴のテキストデータから、ターゲットとなる事象、例えばサービス解約や会員退会の申し出の有無に影響を与える問い合わせ内容（トピック）を抽出すること、あるいは技術動向を探るために、特許文献に代表される技術文書中のテキストデータから、ターゲットとなる事象、例えば出願年に影響を与える技術内容（トピック）を抽出することなどは重要な課題である。

【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が、上記したターゲットとなる事象に着目したものではない。このため、事象の発生の有無に影響を与えるトピックも、そうでないトピックも同様に抽出してしまい、テキストデータに潜む事象に対する要因関係を顕在化できていない。

【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、事象の発生の有無に影響を与えるトピックを優先的に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。

10

【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及び前記テキストデータに関するメタ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個数を表す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、前記メタデータが予め設定した事象に該当するとき、当該メタデータに関する前記テキストデータについて第1の共起行列を作成し、前記メタデータが前記事象に該当しないとき、当該メタデータに関する前記テキストデータ、あるいは前記事象の該当有無にかかわらず全ての前記テキストデータについて第2の共起行列を作成し、前記第1の共起行列と前記第2の共起行列との差を計算することで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。

20

【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、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、前記共起行列作成ステップは、前記メタデータが予め設定した事象に該当するとき、当該メタデータに関する前記第1のカテゴリに分類された前記テキスト部、及び当該メタデータに関する前記第2のカテゴリに分類された前記テキスト部について第1の共起行列を作成し、前記メタデータが前記事象に該当しないとき、当該メタデータに関する前記第1のカテゴリに分類された前記テキスト部、及び当該メタデータに関する前記第2のカテゴリに分類された前記テキスト部、あるいは前記事象の該当有無にかかわらず全ての前記テキスト部について第2の共起行列を作成し、前記第1の共起行列と前記第2の共起行列との差を計算することで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。

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、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語群の出現頻度、並びに前記第2条件付確率

40

50

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、前記共起行列作成ステップは、前記メタデータが予め設定した事象に該当するとき、当該メタデータに関する前記文章について第1の共起行列を作成し、前記メタデータが前記事象に該当しないとき、当該メタデータに関する前記文章、あるいは前記事象の該当有無にかかわらず全ての前記文章について第2の共起行列を作成し、前記第1の共起行列と前記第2の共起行列との差を計算することで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。

【0010】

本発明の第4の態様は、第1又は第2の態様に記載の分析方法において、前記共起行列作成ステップでは、前記第1の共起行列に対して補正值を乗じ、前記補正值は、前記第1の共起行列の作成に用いられた前記テキストデータの件数に対する、前記第2の共起行列の作成に用いられた前記テキストデータの件数の比率、あるいは全ての前記テキストデータの件数の比率であることを特徴とする分析方法にある。

10

【0011】

本発明の第5の態様は、第1又は第2の態様に記載の分析方法において、前記共起行列作成ステップでは、前記第2の共起行列に対して補正值を乗じ、前記補正值は、前記第2の共起行列の作成に用いられた前記テキストデータの件数に対する、前記第1の共起行列の作成に用いられた前記テキストデータの件数の比率、あるいは全ての前記テキストデータの件数の比率であることを特徴とする分析方法にある。

20

【0012】

本発明の第6の態様は、第3の態様に記載の分析方法において、前記共起行列作成ステップでは、前記第1の共起行列に対して補正值を乗じ、前記補正值は、前記第1の共起行列の作成に用いられた前記テキストデータの文章数に対する、前記第2の共起行列の作成に用いられた前記テキストデータの文章数の比率、あるいは全ての前記テキストデータの文章数の比率であることを特徴とする分析方法にある。

【0013】

本発明の第7の態様は、第3の態様に記載の分析方法において、前記共起行列作成ステップでは、前記第2の共起行列に対して補正值を乗じ、前記補正值は、前記第2の共起行列の作成に用いられた前記テキストデータの文章数に対する、前記第1の共起行列の作成に用いられた前記テキストデータの文章数の比率、あるいは全ての前記テキストデータの文章数の比率であることを特徴とする分析方法にある。

30

【0014】

本発明の第8の態様は、第1から第7の何れか一つの態様に記載の分析方法であって、前記トピックごとに、前記スコアが所定の閾値以上である条件の下で前記メタデータが前記事象に該当する確率を前記メタデータが前記事象に該当する確率で除した事象該当ありの指標値、及び前記スコアが所定の閾値以上である条件の下で前記メタデータが前記事象に該当しない確率を前記メタデータが前記事象に該当しない確率で除した事象該当なしの指標値を計算する集計ステップを備えることを特徴とする分析方法にある。

【0015】

40

本発明の第9の態様は、テキストデータ、及び前記テキストデータに関するメタ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個数を表す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

50

手段は、前記メタデータが予め設定した事象に該当するとき、当該メタデータに関する前記テキストデータについて第1の共起行列を作成し、前記メタデータが前記事象に該当しないとき、当該メタデータに関する前記テキストデータ、あるいは前記事象の該当有無にかかわらず全ての前記テキストデータについて第2の共起行列を作成し、前記第1の共起行列と前記第2の共起行列との差を計算することで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。

【0016】

本発明の第10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メタ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個数を表す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、前記メタデータが予め設定した事象に該当するとき、当該メタデータに関する前記テキストデータについて第1の共起行列を作成し、前記メタデータが前記事象に該当しないとき、当該メタデータに関する前記テキストデータ、あるいは前記事象の該当有無にかかわらず全ての前記テキストデータについて第2の共起行列を作成し、前記第1の共起行列と前記第2の共起行列との差を計算することで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】

【0017】

本発明によれば、事象の発生の有無に影響を与えるトピックを優先的に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】

【0018】

【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】トピックのトレンドを示す図である。

【図4】トピックのトレンドを示す図である。

【図5】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】

【0019】

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。

【0020】

〈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。

【0021】

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C P Uにより実行される。

【0022】

このような分析プログラムは、テキストデータ及びメタ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。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。テキストデータの符号化の方式（文字コード）は特に限定ではなく、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。メタデータとは、テキストデータに関連するデータ、あるいはテキストデータから作成したデータである。

【0023】

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「風」及び「空気」を含む10年分（出願日が2006年1月1日から2015年12月31日）の特許出願（30,039件）を抽出し、その特許出願の要約書のうち「解決手段」に記載された文章をテキストデータとする。また、テキストデータに関連するメタデータとして、上記特許出願の公報に記載された書誌事項を用いる。表1にテキストデータ及びメタデータの一例を示す。

【0024】

【表1】

テキストデータID	テキストデータ	メタデータ		
		出願人	発明者	出願年
1	…送風口を有し、前記送風口へ空気を送る送風手段と、送風手段の制御をする制御装置とを備える換気装置。制御手段により、不要な送風を停止することでエネルギー効率が向上した。	A	X	2010
2	微粉炭機で製造された微粉炭を空気とともに燃焼させる燃焼ボイラを備えた石炭火力発電所において、排ガスを循環させる通風機を設け…800°C以上で燃焼する。これにより、溶融スラグの発生が抑えられる。	B	Y	2009
3	空気中の微粒子を捕集する捕集部と、捕集部へ向けて風を送る送風機とを備えた空气净化装置。送風する流量・向きを変化させることで微粒子の捕集性能が向上した。	C	Z	2015

【0025】

表1には、3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。メタデータは、一例として、出願人、発明者、出願年を例示してある。これらのテキストデータ及びメタデータは、電子化された特許公報から得ることができる。

【0026】

このようなテキストデータ及びメタ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、スコア計算手段13、及び集計手段14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11～14として機能させるプログラムである。

【0027】

共起行列作成手段11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。

【0028】

共起行列の作成方法について説明する。まず、共起行列作成手段11は、テキストデータのメタデータが事象に該当するか否かを判定する。事象とは、テキストデータを2つに

分類するために、メタデータに適用される条件である。この事象（条件）に該当したメタデータに関するテキストデータは、後述する第1の共起行列を作成するための入力データとなる。一方、この事象に該当しなかったメタデータに関するテキストデータ、あるいは事象の該当有無にかかわらない全てのテキストデータは、後述する第2の共起行列を作成するための入力データとなる。

【0029】

事象として「出願年は2013年以後である」を例に取り説明する。表1のテキストデータID「3」については、出願年が2015年である。したがって、テキストデータID「3」のメタデータ（出願年）は当該事象に該当する。このようにメタデータが事象に該当したテキストデータを、「事象が発生したテキストデータ」とも称する。

10

【0030】

一方、表1のテキストデータID「1」「2」については、出願年が2010、2009年である。したがって、テキストデータID「1」「2」のメタデータ（出願年）は当該事象に該当しない。このようにメタデータが事象に該当しないテキストデータを、「事象が発生しなかったテキストデータ」とも称する。

10

【0031】

このような事象は、予めプログラムなどに設定しておく。事象の選び方には特に限定はない。上述した事象は、出願年が2013年を境にして、後述するトピックに変化があるかを分析するという目的で定めたものである。

20

【0032】

共起行列作成手段11は、事象が発生したテキストデータについて第1の共起行列を作成し、かつ、事象が発生しなかったテキストデータについて、あるいは事象の発生有無にかかわらない全てのテキストデータについて第2の共起行列を作成する。

20

【0033】

第1の共起行列及び第2の共起行列は、入力データが異なるだけであり、具体的な作成ステップは同じであるから、第1の共起行列を例にとり説明する。

30

【0034】

テキストデータには、複数の文章が含まれることがある。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。分析装置1で実行される分析プログラム10の共起行列作成手段11は、テキストデータを一つずつ読み込み、各テキストデータについて、句点や「?」「!」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、次のように2つの文章が抽出される。

30

【0035】

【表2】

テキストデータID	文章ID	テキストデータ
1	1	…送風口を有し、前記送風口へ空気を送る送風手段と、送風手段の制御をする制御装置とを備える換気装置。
1	2	制御手段により、不要な送風を停止することでエネルギー効率が向上した。

40

【0036】

文章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、一つの文章IDについては、表1に示したメタデータも関連づけられていることになる。

【0037】

一つのテキストデータは、発明を特定する事項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で記載されていることが多い。表2のテキストデータID「1」からは、換気装置の構成について述べた文章（文章ID「1」）、及び換気装置の効果について述べた文章（文章ID「2」）が得られることになる。

【0038】

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

50

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。

【0039】

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語群及び第2語群を抽出する。第1語群及び第2語群は、文章中に含まれる特定の品詞に分類される単語や、係り受け表現（文法的構造を持つ単語と単語のペア）からなる。第1語群と第2語群とで、異なる語群が抽出されるようにする。例えば、文章から「単語」を抽出し、その結果を第1語群とし、文章から「係り受け表現」を抽出し、その結果を第2語群とする。

10

【0040】

もちろん、第1語群と第2語群の単語等の選び方は特に限定はない。例えば、文章中に含まれる単語のうち「名詞」で分類される単語を第1語群に、「動詞および形容詞」で分類される単語を第2語群としてもよい。すなわち複数の品詞を用いて第1語群（又は第2語群）を形成してもよい。

【0041】

共起行列作成手段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あるいは構文解析手法を適用することで、一つの文章の中から第1語群及び第2語群を抽出する。

20

【0042】

そして、共起行列作成手段11は、文章より抽出された第1語群及び第2語群から、共起行列を集計する。共起行列とは、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の個数を表したものである。表3に第1の共起行列（一部）、表4に第2の共起行列（一部）を例示する。以下例では、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。

【0043】

【表3】

	空気-吸い込む	吸い込む-空気	連-通す	備える-構成	空気-吹き出す	吹出口-吹き出す	空気-供給	空気-送風	空気-排出	制御部-備える
配置	66	52	41	43	42	44	23	41	30	24
供給	32	36	14	27	21	13	170	16	21	27
内部	41	43	34	20	29	24	25	22	34	13
送風機	80	63	33	19	39	29	34	38	18	24
制御	27	22	8	17	25	33	17	35	14	122
位置	28	26	16	14	25	22	10	18	10	16
吸い込む	332	285	27	23	120	98	19	25	28	15
吹出口	100	116	26	22	123	215	12	34	9	18
発生	36	35	13	22	21	19	12	10	12	18
外部	38	24	30	19	25	16	13	19	50	12

【0044】

【表4】

40

	空気-吸い込む	吸い込む-空気	連-通す	備える-構成	空気-吹き出す	吹出口-吹き出す	空気-供給	空気-送風	空気-排出	制御部-備える
配置	144	139	109	102	81	99	77	86	70	35
供給	95	65	53	81	33	40	505	57	60	38
内部	93	64	92	44	47	20	59	49	77	26
送風機	145	134	79	76	74	44	91	71	40	48
制御	54	54	21	47	47	70	51	57	33	180
位置	54	46	54	58	43	50	29	26	41	12
吸い込む	814	632	72	52	193	146	49	45	84	43
吹出口	167	185	81	86	247	480	30	29	45	42
発生	82	67	50	45	59	45	37	32	28	29
外部	104	75	63	55	45	23	49	46	119	24

【0045】

第1語群に属する単語として「配置」「供給」「内部」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「空気-吸い込む」「吸い込む-空気」「連-通す」などが列方向に並んでいる。共起行列作成手段11は、一つの文章の中に、「配置」と「空気-吸い込む」との組み合わせが存在すれば、一つカウントする。この組み合わせを共起ペアと称する。表3の第1の共起行列の例では、「配置」及び「空気-吸い込む」という共起ペアが一つの文章の中に存在する文章数は66件あることになる。

【0046】

次に、共起行列作成手段11は、第1の共起行列と第2の共起行列の差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。ここでいう第1の共起行列と第2の共起行列の差とは、第1語群に属する語と、第2語群に属する語が同じものについて、組み合わせ数の差をいう。表3、表4の例では、第1の共起行列及び第2の共起行列の同じ共起ペア同士の差を取り。差の取り方は、差の絶対値としてもよいし、差の二乗としてもよい。いずれにしても差が負にならないようにする。

10

【0047】

また、第1の共起行列を得るために用いられたテキストデータの文章数（事象が発生したテキストデータを構成する文章の数）と、第2の共起行列を得るために用いられたテキストデータの文章数（事象が発生しなかったテキストデータを構成する文章の数）とに差がある場合は、第1の共起行列又は第2の共起行列の一方あるいはその両方を補正することが好みしい。

20

【0048】

本実施形態で示す例では、第1の共起行列を得るために用いたテキストデータの文章数が11, 831件、第2の共起行列を得るために用いたテキストデータの文章数が33, 283件であった。この場合、第2の共起行列の全ての共起ペアの個数に、11, 831 / 33, 283 (≈ 0.3555) を乗じる補正を行う。つまり、文章数の多いテキストデータから作成された第2の共起行列に対して、その文章数(33, 283件)に対する、第1の共起行列の作成に用いられたテキストデータの文章数(11, 831件)の比率を補正值とする。表5に、第1の共起行列と、上述したような補正をした第2の共起行列との差の絶対値を取った共起行列(一部)を示す。

30

【0049】**【表5】**

	空気-吸い込む	吸い込む-空気	連-通す	備える-構成	空気-吹き出す	吹出口-吹き出す	空気-供給	空気-送風	空気-排出	制御部-備える
配置	14.8	2.6	2.3	6.7	13.2	8.8	4.4	10.4	5.1	11.6
供給	1.8	12.9	4.8	1.8	9.3	1.2	9.5	4.3	0.3	13.5
内部	7.9	20.3	1.3	4.4	12.3	16.9	4.0	4.6	6.6	3.8
送風機	28.5	15.4	4.9	8.0	12.7	13.4	1.7	12.8	3.8	6.9
制御	7.8	2.8	0.5	0.3	8.3	8.1	1.1	14.7	2.3	58.0
位置	8.8	9.6	3.2	6.6	9.7	4.2	0.3	8.8	4.6	11.7
吸い込む	42.7	60.3	1.4	4.5	51.4	46.1	1.6	9.0	1.9	0.3
吹出口	40.6	50.2	2.8	8.6	35.2	44.4	1.3	23.7	7.0	3.1
発生	6.9	11.2	4.8	6.0	0.0	3.0	1.2	1.4	2.0	7.7
外部	1.0	2.7	7.6	0.6	9.0	7.8	4.4	2.6	7.7	3.5

【0050】

このような補正を行うことで、第2の共起行列は、第1の共起行列と同じ11, 831件のテキストデータの文章を用いて作成したものと同等と考えられる。このような補正是、第1の共起行列と第2の共起行列のテキストデータの文章数に偏りがある場合に特に有用である。

【0051】

なお、第1の共起行列の共起ペアに補正值を乗じてもよい。この場合は、第1の共起行列の作成に用いられたテキストデータの文章数(11, 831件)に対する、第2の共起行列の作成に用いられたテキストデータの文章数(33, 283件)の比率を補正值とす

る。また、第1の共起行列の共起ペアと第2の共起行列の共起ペアの両方に補正值を乗じてもよい。この場合は、第1の共起行列の共起ペアには、第1の共起行列の作成に用いられたテキストデータの文章数（11, 831件）に対する、全テキストデータの文章総数（45, 114件）の比率を補正值とし、第2の共起行列の共起ペアには、第2の共起行列の作成に用いられたテキストデータの文章数（33, 283件）に対する、全テキストデータの文章総数（45, 114件）の比率を補正值とする。

【0052】

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。

10

【0053】

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。

20

【0054】

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T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T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Tに属する文章Dの条件付確率及び各トピックTに属する単語Wの条件付確率及びトピックTの確率がEMアルゴリズムにより計算される。

20

【0055】

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T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T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T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 $P(W_1 | T)$ 、及び各トピックT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 $P(W_2 | T)$ 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。

30

【0056】

表6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6には、複数作成されたトピックのうち、2つのトピックT10とトピックT13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いる。

40

【0057】

【表 6】

トピックT10				トピックT13			
確率	単語	確率	係り受け	確率	単語	確率	係り受け
3.4%	塵埃	0.9%	付着-塵埃	1.5%	配置	0.5%	流れる-空気
1.8%	送風機	0.8%	塵埃-除去	1.1%	流れる	0.4%	空気-流す
1.7%	掃除機	0.8%	吸い込む-塵埃	1.0%	向ける	0.4%	備える-車両用空調装置
1.6%	吸い込む	0.7%	塵埃-含む	1.0%	下流	0.3%	空気-送風
1.3%	分離	0.7%	塵埃-吸い込む	1.0%	車両用空調装置	0.3%	下流-配置
1.3%	フィルタ	0.7%	塵埃-分離	0.9%	方向	0.3%	通過-空気
1.3%	捕集	0.6%	発生-送風機	0.9%	車室内	0.3%	前方-配置
1.0%	集塵部	0.6%	含む-空気	0.9%	車両	0.3%	方向-沿う
...

【0058】

トピックT10についてみると、第1条件付確率が最上位である単語は「塵埃」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「付着-塵埃」である。このようなトピックT10に所属する単語及び係り受け表現に基づいて、トピックT10の意味を解釈することができる。例えば、トピックT10は、第1条件付確率が上位である単語に基づけば、塵埃の分離に関するトピックであると解釈することができる。

【0059】

20

PLSAは、トピック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。

【0060】

30

本実施形態では、表7に示すように、トピック抽出手段12により14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7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。

【0061】

【表7】

No.	トピック名	No.	トピック名
T01	冷凍サイクル	T08	イオン発生
T02	空気の冷却	T09	電解水の生成
T03	冷却ファン	T10	塵埃の分離
T04	空気流(吸込と吹出)	T11	羽の回転
T05	紙葉類の搬送	T12	検出と制御
T06	衣類乾燥	T13	車両用空調の配置
T07	空気の燃焼	T14	配置と形成

【0062】

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

50

テキストデータに対する各トピックのスコアを求める。

【0063】

各トピック T_k を条件とした各文章 S_h の条件付確率を $P(S_h | T_k)$ とする。各文章の発生確率を $P(S_h)$ とする。各文章に対するトピックのスコアは、 $P(S_h | T_k) / P(S_h)$ である。スコア計算手段13は、 $P(S_h | T_k)$ 及び $P(S_h)$ を次のように計算する。なお、 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。

【0064】

【数1】

10

$$(1) \quad P(S_{W_h} | T_k) = \sum_i P(S_{W_h} | W_i) P(W_i | T_k)$$

$$(2) \quad P(S_{E_h} | T_k) = \sum_j P(S_{E_h} | E_j) P(E_j | T_k)$$

$$(3) \quad P(S_h | T_k) = P(S_h | S_{W_h}) P(S_{W_h} | T_k) + P(S_h | S_{E_h}) P(S_{E_h} | T_k)$$

$$(4) \quad P(S_h) = \sum_k P(S_h | T_k) P(T_k)$$

【0065】

各文章 S_h について、第1語群に設定した単語 W によって定義される文章を S_{Wh} 、第2語群に設定した係り受け表現 E によって定義される文章を S_{eh} とする。 $P(S_h | T_k)$ を計算するにあたり、 $P(S_{Wh} | T_k)$ と $P(S_{eh} | T_k)$ を計算する。これらはそれぞれ上記式(1) (2) で計算される。単語 W_i が含まれる文章の数を $n(W_i)$ 、係り受け表現 E_j が含まれる文章の数を $n(E_j)$ とすると、 $P(S_{Wh} | W_i)$ は $n(W_i)$ の逆数、 $P(S_{eh} | E_j)$ は $n(E_j)$ の逆数として計算される。 $P(W_i | T_k)$ と $P(E_j | T_k)$ は、PLSAの実行によって得られる第1条件付確率と第2条件付確率である。

【0066】

30

$P(S_h | T_k)$ は、上記式(3)より得られる。 $P(S_h | S_{Wh})$ と $P(S_h | S_{eh})$ は文章 S_h において重みは同じといえるので、それぞれ0.5とする。 $P(S_h)$ は、上記式(4)で計算され、 $P(T_k)$ はPLSAの実行により得られる。

【0067】

上記式(3)の $P(S_h | T_k)$ を、上記式(4)の $P(S_h)$ で除した値が各文章のスコアとなる。本実施形態では、各文章の発生確率を上記式(4)のように計算しているが、例えば一様分布に従うと仮定し、 $P(S_h)$ を文章の総数の逆数とするなど、各文章の発生確率の取り方はこれに限らない。

【0068】

40

このように、 $P(S_h | T_k)$ と $P(S_h)$ との比をもって文章 S_h におけるトピック T_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 T_k を条件とすることで上昇し、トピック T_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 T_k の関係の強さを把握しやすくなることができる。表8に各文章 S_h に対する各トピック T_k のスコアを例示する。

【0069】

【表 8】

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

【0070】

10

例えば、文章ID「1」は、トピックT1についてのスコアが3.1であり、トピックT2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ている。

【0071】

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。

【0072】

表9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。

20

【0073】

【表9】

テキストデータID	文章ID(h)	トピックTk			
		T1	T2	...	T14
1	1	3.1	0.9		1.1
1	2	1.4	0.2		2.4
2	3	0.8	5.8		0.9
2	4	1.2	3.2		1.0
2	5	0.6	1.8		1.6

【0074】

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トピックT1～T14のスコアについて、トピック毎に最大値（文章「1」と文章「2」のうち大きいスコア）を求める。

【0075】

文章「1」に対するトピックT1のスコアは「3.1」であり、文章「2」に対するトピックT1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「3.1」がテキストデータ「1」に対するトピックT1のスコアとなる。以下同様に、トピックT2～T14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9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。

40

【0076】

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10にフラグ情報を示す。

【0077】

【表10】

テキスト データID	トピックTk			
	T1	T2	...	T14
1	1	0		0
2	0	1		0

【0078】

テキストデータ「1」は、トピックT1のスコアが「3.1」であるから（表9参照）
、フラグ情報は「1」となり、トピックT2のスコアは「0.9」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。 $P(S_h | T_k) / P(S_h)$ 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。

10

【0079】

次に、上述したスコアに基づいて、事象の発生の有無によってテキストデータ（特許出願）の件数がどの様に変化するかを集計することについて説明する。

【0080】

まず、集計手段14は、事象Xの発生の有無とトピックTの関連度を示す指標値として「事象該当ありの指標値」及び「事象該当なしの指標値」を計算する。なおメタデータが事象Xに該当する場合はX=1、該当しない場合はX=0と表記する。

20

【0081】

「事象該当ありの指標値」は、トピックTのスコアが所定の閾値以上（T=1と表記する）である条件の下でメタデータが事象Xに該当する確率を、メタデータが事象Xに該当する確率で除した値である。

【0082】

「トピックTのスコアが所定の閾値以上である条件の下でメタデータが事象Xに該当する確率」を $P(X=1 | T=1)$ と表記する。また、「メタデータが事象Xに該当する確率」を $P(X=1)$ と表記する。これらを用いると、「事象該当ありの指標値」は、 $P(X=1 | T=1) / P(X=1)$ で求められる。

【0083】

30

本実施形態で取り上げる事象は「出願年が2013年以後」であるから、 $P(X=1 | T=1)$ は、トピックTのスコアが所定の閾値以上である条件の下で、出願年が2013年以後である確率を表している。

【0084】

なお、 $P(X=1)$ は、出願年が2013年以後であるテキストデータの件数をテキストデータの総数で除することで得られる。また、 $P(X=1 | T=1)$ は、あるトピックTについてのフラグ情報が「1」であるテキストデータの件数のうち、出願年が2013年以後であるテキストデータの件数が占める割合を求めることで得られる。

【0085】

「事象該当なしの指標値」は、トピックTのスコアが所定の閾値以上である条件の下でメタデータが事象Xに該当しない確率を、メタデータが事象Xに該当しない確率で除した値である。

40

【0086】

「トピックTのスコアが所定の閾値以上である条件の下でメタデータが事象Xに該当しない確率」を $P(X=0 | T=1)$ と表記する。また、「メタデータが事象Xに該当しない確率」を $P(X=0)$ と表記する。これらを用いると、「事象該当なしの指標値」は、 $P(X=0 | T=1) / P(X=0)$ で求められる。

【0087】

本実施形態で取り上げる事象は「出願年が2013年以後」であるから、 $P(X=0 | T=1)$ は、トピックTのスコアが所定の閾値以上である条件の下で、出願年が2012

50

年以前である確率を表している。

【0088】

なお、 $P(X=0)$ は、出願年が 2012 年以前であるテキストデータの件数をテキストデータの総数で除することで得られる。また、 $P(X=0 | T=1)$ は、あるトピック T についてのフラグ情報が「1」であるテキストデータの件数のうち、出願年が 2012 年以前であるテキストデータの件数が占める割合を求めて得られる。

【0089】

集計手段 14 は、上記した事象該当ありの指標値と事象該当なしの指標値をトピック毎に計算し、本実施形態の例ではそれらの比率（増減率）を計算する。ここでは、事象該当なしの指標値に対する事象該当ありの指標値の比率（事象該当ありの指標値／事象該当なしの指標値）を計算した。これは、あるトピックが 2013 年の前後においてどの程度増えたか、又は減ったかを 2012 年以前と 2013 年以後の各テキストデータの件数を加味して示すものとなる。このような比率の例を表 11 に示す。なお、減った場合はマイナスを付している。

【0090】

【表 11】

増減率	トピック名	増減率	トピック名
48.5%	T13.車両用空調の配置	7.0%	T01.冷凍サイクル
30.3%	T14.配置と形成	-0.3%	T05.紙葉類の搬送
23.9%	T04.空気流(吸込と吹出)	-5.9%	T07.空気の燃焼
17.7%	T12.検出と制御	-8.8%	T09.電解水の生成
14.9%	T11.羽の回転	-17.0%	T10.塵埃の分離
13.1%	T03.冷却ファン	-24.2%	T06.衣類乾燥
11.6%	T02.空気の冷却	-28.7%	T08.イオン発生

【0091】

トピック T13 は、2012 年以前と比較して、2013 年以後では指標値が 48.5% 増加し、トピック T08 は、2012 年以前と比較して、2013 年以後では指標値が 28.7% 減少している。

【0092】

本実施形態の例では、集計手段 14 で計算した事象該当ありの指標値と事象該当なしの指標値について、2013 年以後と 2012 年以前とに分けてその比率（増減率）を計算したが、このような態様に限定されない。例えば、図 3、図 4 に示すように、年毎に上記指標値を並べ、各トピックのトレンドを表示するようにしてもよい。

【0093】

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 5 は、分析装置での処理を示すフローチャートである。

【0094】

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する。具体例については、上述したので説明は省略する。

【0095】

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。

【0096】

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。

【0097】

10

次に、スコアに基づいて、トピックごとに事象該当ありの指標値と事象該当なしの指標値を計算する（ステップS4：集計ステップ）。具体例については上述したので説明は省略する。

【0098】

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、事象が発生したテキストデータと事象が発生していないテキストデータに分け、第1の共起行列及び第2の共起行列を作成し、これらの差から得られたものである。

【0099】

20

このようにして得られた共起行列は、事象の発生有無に関係する共起ペアは頻度が大きくなり、そうでない共起ペアでは頻度が小さくなる。このため、共起行列にPLSAを適用する結果、事象の発生有無に影響を与えるトピックを優先的に抽出することができる。このように、本発明によれば、テキストデータに潜む要因関係（トピックと事象との関係）を顕在化することができる。

【0100】

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。

30

【0101】

本発明では、事象該当ありの指標値と事象該当なしの指標値をトピック毎に計算する。これにより、各トピックが着目する事象の有無（上記実施例では2013年前後の出願年であるか否か）に対してどの程度関連するものなのかを明確にできる。

【0102】

本発明では、第1の共起行列と第2の共起行列との差を取る際に、何れか一方あるいは両方に補正值を乗じて補正した。これにより、事象の発生有無に影響を与えるトピックを、その発生有無のデータ規模によらず抽出することができる。

【0103】

40

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11～14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。

【0104】

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、商品の顧客満足度（ターゲットとなる事象）に影響を与えるトピックを当該テキストデータから抽出するなど、テキストデータの一般に適用することができる。

【0105】

〈比較例〉

50

上述した実施形態と同じテキストデータを用いて、第1の共起行列及び第2の共起行列を作成せずに、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。つまり、共起行列の作成方法自体は、第1の共起行列及び第2の共起行列と同様であり、テキストデータのメタデータが事象に該当するか否かの判定を行わずに、全てのテキストデータを入力データとした点が異なる。

【0106】

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12に示す。本発明では表7に示したように、14個のトピックが抽出されたが、比較例においては47個のトピックが抽出された。10

【0107】

【表12】

No.	トピック名	No.	トピック名
t01	冷凍サイクル	t25	加湿
t02	冷却	t26	放電式ミスト生成
t03	車室内空調	t27	微細粒子の飛散(マイナスイオン等)
t04	空気路	t28	イオン発生・空気除菌・脱臭
t05	換気	t29	電解水生成と除菌
t06	排気	t30	空气净化 & 効率性
t07	空気の吸込と吹出	t31	塵埃除去
t08	流体の流入と吐出	t32	塵埃分離
t09	空気流の利用と制御	t33	回転駆動
t10	送風	t34	電源と駆動制御
t11	空気の噴出	t35	運転と停止の制御
t12	送風搬送(紙類類等)	t36	センサと制御(温度や風量等)
t13	印刷	t37	人検出
t14	光の利用(照射、発光等)	t38	風向制御
t15	ファンと機器冷却	t39	抑制・防止(騒音やコスト等)
t16	空気導入と車両エンジンの冷却	t40	構成・取り付け
t17	放熱	t41	接続
t18	除湿	t42	機器(熱交換器等)の配置
t19	乾燥機能	t43	配置と形成
t20	洗濯乾燥	t44	位置・形状・大きさ
t21	洗净(衣類や食器等)	t45	位置の方向
t22	燃焼	t46	方法・装置
t23	加熱	t47	その他(発明目的、ケース構成等)
t24	温湿度制御と空気循環		

【0108】

さらに、得られたトピックについて上述した実施形態と同様にスコア集計した結果を表13に示す。47個のトピックについて、2013年前後における指標値の増減率が得られた。40

【0109】

【表13】

増減率	トピック名	増減率	トピック名
19.3%	t44.位置・形状・大きさ	0.0%	t06.排気
17.3%	t09.空気流の利用と制御	-1.1%	t46.方法・装置
16.1%	t43.配置と形成	-1.3%	t22.燃焼
15.0%	t36.センサと制御(温度や風量等)	-1.6%	t23.加熱
14.0%	t38.風向制御	-2.2%	t02.冷却
13.5%	t16.空気導入と車両エンジンの冷却	-2.3%	t47.その他(発明目的、ケース構成等)
12.9%	t14.光の利用(照射、発光等)	-3.8%	t21.洗浄(衣類や食器等)
12.2%	t45.位置の方向	-4.4%	t13.印刷
11.3%	t10.送風	-4.8%	t05.換気
10.6%	t03.車室内空調	-5.0%	t15.ファンと機器冷却
9.6%	t37.人検出	-6.4%	t12.送風搬送(紙類類等)
8.9%	t41.接続	-7.3%	t29.電解水生成と除菌
7.9%	t40.構成・取り付け	-9.8%	t32.塵埃分離
7.4%	t07.空気の吸込と吹出	-9.9%	t27.微細粒子の飛散(マイナスイオン等)
6.5%	t33.回転駆動	-15.5%	t24.温湿度制御と空気循環
6.3%	t34.電源と駆動制御	-15.7%	t30.空気浄化&効率性
3.9%	t04.空気路	-17.9%	t20.洗濯乾燥
3.7%	t17.放熱	-19.5%	t31.塵埃除去
3.3%	t01.冷凍サイクル	-20.7%	t28.イオン発生・空気除菌・脱臭
3.1%	t08.流体の流入と吐出	-21.3%	t39.抑制・防止(騒音やコスト等)
2.5%	t25.加湿	-22.0%	t19.乾燥機能
1.9%	t11.空気の噴出	-24.0%	t18.除湿
0.8%	t35.運転と停止の制御	-27.4%	t26.放電式ミスト生成
0.4%	t42.機器(熱交換器等)の配置		

【0110】

比較例においては、47個のトピックは、2013年前後の増減率がばらついていることが分かる。一方、表11に示すように、本発明によれば、2013年前後の増減率は高いものと低いものに集中しており、その値も表13よりも高く、事象（2013年前後ににおける出願傾向）に影響を与える14個のトピックが優先的に抽出されている。

【0111】

30

<実施形態2>

実施形態1では、テキストデータからそこに含まれる文章を抽出し、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、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。

【0112】

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個数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。

40

【0113】

共起行列の作成方法について説明する。まず、共起行列作成手段11は、テキストデータのメタデータが事象に該当するか否かを判定する。この判定については、実施形態1で説明したので、ここでの説明は省略する。

【0114】

共起行列作成手段11は、事象が発生した全てのテキストデータから第1語群及び第2語群を抽出する。そして、共起行列作成手段11は、抽出された第1語群及び第2語群から第1の共起行列を集計する。

【0115】

同様に、共起行列作成手段11は、事象が発生しなかった全てのテキストデータ（また

50

は事象の発生有無にかかわらない全てのテキストデータ）から第1語群及び第2語群を抽出する。そして、共起行列作成手段11は、抽出された第1語群及び第2語群から第2の共起行列を集計する。

【0116】

このようにして、第1語群及び第2語群に属する具体的な語や件数は異なるが、表3及び表4のような第1の共起行列及び第2の共起行列が得られる。表3がテキストデータから作成された第1の共起行列であると仮定すると、「配置」及び「空気一吸い込む」という共起ペアが存在するテキストデータの数は66件であることを表す。

【0117】

また、第1の共起行列を得るために用いられたテキストデータの数（事象が発生したテキストデータの数）と、第2の共起行列を得るために用いられたテキストデータの数（事象が発生しなかったテキストデータの数）とに差がある場合は、第1の共起行列又は第2の共起行列の一方あるいはその両方を補正することが好ましい。

10

【0118】

例えば、第1の共起行列を得るために用いたテキストデータの数がN1件、第2の共起行列を得るために用いたテキストデータの数がN2件であった。N1 < N2とする。この場合、第2の共起行列の全ての共起ペアの個数に、N1 / N2を乗じる補正を行う。つまり、数の多いテキストデータから作成された第2の共起行列に対して、第2の共起行列の作成に用いられたテキストデータの数（N2）に対する、第1の共起行列の作成に用いられたテキストデータの数（N1）の比率を補正值とする。もちろん、第1の共起行列を補正してもよい。この場合は、第1の共起行列の全ての共起ペアの個数に、補正值N2 / N1を乗じる。

20

【0119】

このような補正を行うことで、第2の共起行列は、第1の共起行列と同じN1件のテキストデータを用いて作成したものと同等と考えられる。このような補正是、第1の共起行列と第2の共起行列のテキストデータの数に偏りがある場合に特に有用である。

【0120】

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。

30

【0121】

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。

【0122】

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。

30

【0123】

各トピックTkを条件とした各テキストデータShの条件付確率をP(Sh | Tk)とする。各テキストデータの発生確率をP(Sh)とする。各テキストデータに対するトピックのスコアは、 $P(Sh | Tk) / P(Sh)$ である。スコア計算手段13は、 $P(Sh | Tk)$ 及び $P(Sh)$ を次のように計算する。なお、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。

40

【0124】

【数2】

- $$(1) P(S_{W_h}|T_k) = \sum_i P(S_{W_h}|W_i) P(W_i|T_k)$$
- $$(2) P(S_{e_h}|T_k) = \sum_j P(S_{e_h}|E_j) P(E_j|T_k)$$
- $$(3) P(S_h|T_k) = P(S_h|S_{W_h}) P(S_{W_h}|T_k) + P(S_h|S_{e_h}) P(S_{e_h}|T_k)$$
- $$(4) P(S_h) = \sum_k P(S_h|T_k) P(T_k)$$

【0125】

各テキストデータ S_h について、第1語群に設定した単語 W によって定義されるテキストデータを S_{wh} 、第2語群に設定した係り受け表現 E によって定義されるテキストデータを S_{eh} とする。 $P(S_h | T_k)$ を計算するにあたり、 $P(S_{wh} | T_k)$ と $P(S_{eh} | T_k)$ を計算する。これらはそれぞれ上記式(1) (2)で計算される。単語 W_i が含まれるテキストデータの数を $n(W_i)$ 、係り受け表現 E_j が含まれるテキストデータの数を $n(E_j)$ とすると、 $P(S_{wh} | W_i)$ は $n(W_i)$ の逆数、 $P(S_{eh} | E_j)$ は $n(E_j)$ の逆数として計算される。 $P(W_i | T_k)$ と $P(E_j | T_k)$ は、PLSAの実行によって得られる第1条件付確率と第2条件付確率である。

20

【0126】

$P(S_h | T_k)$ は、上記式(3)より得られる。 $P(S_h | S_{wh})$ と $P(S_h | S_{eh})$ は文章 S_h において重みは同じといえるので、それぞれ0.5とする。 $P(S_h)$ は、上記式(4)で計算され、 $P(T_k)$ はPLSAの実行により得られる。

【0127】

上記式(3)の $P(S_h | T_k)$ を、上記式(4)の $P(S_h)$ で除した値が各テキストデータのスコアとなる。本実施形態では、各テキストデータの発生確率を上記式(4)のように計算しているが、例えば一様分布に従うと仮定し、 $P(S_h)$ をテキストデータの総数の逆数とするなど、各テキストデータの発生確率の取り方はこれに限らない。

【0128】

30

このように、 $P(S_h | T_k)$ と $P(S_h)$ の比をもってテキストデータ S_h におけるトピック T_k のスコアとする。この値が1を超えるということは、テキストデータ S_h の発生確率はトピック T_k を条件とすることで上昇し、トピック T_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、各テキストデータ S_h とトピック T_k の関係の強さを把握しやすくすることができる。表14に各テキストデータ S_h に対する各トピック T_k のスコアを例示する。

【0129】

【表14】

テキストデータID	トピックTk			
	T1	T2	...	T14
1	$P(S1 T1)/P(S1)=3.1$	$P(S1 T2)/P(S1)=0.9$		$P(S1 T14)/P(S1)=1.1$
2	$P(S5 T1)/P(S5)=0.6$	$P(S5 T2)/P(S5)=1.8$		$P(S5 T14)/P(S5)=1.6$
...

【0130】

例えば、テキストデータID「1」は、トピックT1についてのスコアが3.1であり、トピックT2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ている。

【0131】

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

50

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」、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15にフラグ情報を示す。

【0132】

【表15】

テキスト データID	トピックTk			
	T1	T2	...	T14
1	1	0		0
2	0	1		0

【0133】

10

テキストデータ「1」は、トピックT1のスコアが「3.1」であるから（表14参照）、フラグ情報は「1」となり、トピックT2のスコアは「0.9」であるから、フラグ情報は「0」となる。

【0134】

集計手段については、上記スコアを元にして、実施形態1と同様に処理することができる、ここでの説明は省略する。

【0135】

20

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。

【0136】

〈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。

【0137】

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16にテキストデータを例示する。

【表16】

30

テキストデータID	カテゴリ	テキスト部	メタデータ		
			出願人	発明者	出願年
1	タイトル	換気装置	A	X	2010
	課題	…不要な送風が生じていたため、エネルギー効率が悪化していた。			
	解決手段	送風口を有し、前記送風口へ空気を送る送風手段と、送風手段の制御をする制御装置とを備える換気装置。			
	効果	制御手段により、不要な送風を停止することでエネルギー効率が向上した。			
2	タイトル	石炭火力発電所	B	Y	2009
	課題	…温度の低下により、溶融スラグが大量に発生していた。			
	解決手段	…微粉炭機で製造された微粉炭を空気とともに燃焼させる燃焼ボイラを備えた石炭火力発電所において、排ガスを循環させる通風機を設け、さらに、燃焼ボイラにて800°C以上で微粉炭を燃焼する。			
	効果	溶融スラグの発生が抑えられる。			

【0138】

表16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

50

、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。

【0139】

共起行列作成手段11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。

【0140】

共起行列作成手段11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個数を表す共起行列を作成する。

10

【0141】

具体的には、まず、共起行列作成手段11は、テキストデータのメタデータが事象に該当するか否かを判定する。この判定については、実施形態1で説明したので、ここでの説明は省略する。

【0142】

共起行列作成手段11は、事象が発生した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。そして、共起行列作成手段11は、抽出された第1語群及び第2語群から第1の共起行列を集計する。

20

【0143】

同様に、共起行列作成手段11は、事象が発生しなかった全てのテキストデータ（または事象の発生有無にかかわらない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。そして、共起行列作成手段11は、抽出された第1語群及び第2語群から第2の共起行列を集計する。

【0144】

【表17】

		解決手段									
		空気-吸い込む	吸い込む-空気	連-通す	備える-構成	空気-吹き出す	吹出口-吹き出す	空気-供給	空気-送風	空気-排出	制御部-備える
タイトル	燃焼	66	52	41	43	42	44	23	41	30	24
	供給	32	36	14	27	21	13	170	16	21	27
	送風機	41	43	34	20	29	24	25	22	34	13
	制御	80	63	33	19	39	29	34	38	18	24
	吸引	27	22	8	17	25	33	17	35	14	122
	石炭	28	26	16	14	25	22	10	18	10	16
	発電	332	285	27	23	120	98	19	25	28	15
	空気	100	116	26	22	123	215	12	34	9	18
	循環	36	35	13	22	21	19	12	10	12	18
	装置	38	24	30	19	25	16	13	19	50	12

【0145】

表17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」として作成した第1の共起行列を例示している。

40

【0146】

例えば、第1のカテゴリ「タイトル」に分類されたテキスト部に「燃焼」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「空気-吸い込む」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66件であることを表す。第2の共起行列については特に例示しないが、表17と同様の結果が得られる。

【0147】

また、第1の共起行列を得るために用いられたテキストデータの数（事象が発生したテキストデータの数）と、第2の共起行列を得るために用いられたテキストデータの数（事

50

象が発生しなかったテキストデータの数) とに差がある場合は、第1の共起行列又は第2の共起行列の一方あるいはその両方を補正することが好ましい。補正の方法は、実施形態2と同様である。

【0148】

以後の処理は実施形態2と同様である。具体的には、本実施形態の分析方法、分析装置及び分析プログラムは、共起行列作成手段11が第1の共起行列及び第2の共起行列に基づいて共起行列を作成し、トピック抽出手段12がトピックを抽出し、スコア計算手段13がスコアを計算し、集計手段14が集計を行う。

【0149】

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。

10

【0150】

なお、本発明では、メタデータは、事象に該当するか否かによって第1の共起行列及び第2の共起行列を作成するために用いられる。しかしながら、本実施形態のようにカテゴリライズされたテキスト部を用いる場合においては、メタデータをカテゴリライズされたテキスト部として用いてもよい。

【0151】

【表18】

20

テキストデータID	カテゴリ	テキスト部	メタデータ		
			出願人	発明者	出願年
1	タイトル	換気装置	A	X	2010
	課題	…不要な送風が生じていたため、エネルギー効率が悪化していた。			
	解決手段	送風口を有し、前記送風口へ空気を送る送風手段と、送風手段の制御をする制御装置とを備える換気装置。			
	効果	制御手段により、不要な送風を停止することでエネルギー効率が向上した。			
	出願人	A			
	発明者	X			
2	タイトル	石炭火力発電所	B	Y	2009
	課題	…温度の低下により、溶融スラグが大量に発生していた。			
	解決手段	…微粉炭機で製造された微粉炭を空気とともに燃焼させる燃焼ボイラを備えた石炭火力発電所において、排ガスを循環させる通風機を設け、さらに、燃焼ボイラにて800°C以上で微粉炭を燃焼する。			
	効果	溶融スラグの発生が抑えられる。			
	出願人	B			
	発明者	Y			

【0152】

40

表18は、メタデータをカテゴリライズされたテキスト部としても用いる場合のテキストデータの例を示している。表16と同様に、メタデータとして「出願人」「発明者」「出願年」がある。一方、これらの「出願人」「発明者」「出願年」は、テキストデータに含まれるカテゴリとしても用いることができる。出願人の「A」や「B」はカテゴリ「出願人」のテキスト部の具体例である。発明者についても同様である。

【0153】

表19に、第1のカテゴリを「出願人」とし、第2のカテゴリを「解決手段」とし、第1語群を「名詞（人名・法人名）」とし、第2語群を「係り受け表現」とし、表18のテキストデータから作成した第1の共起行列を例示する。

【0154】

50

【表 1 9】

		解決手段									
		空気-吸い込む	吸い込む-空気	連-通す	備える-構成	空気-吹き出す	吹出口-吹き出す	空気-供給	空気-送風	空気-排出	制御部-備える
出願人	A	85	47	119	55	37	129	87	80	76	71
	B	115	122	117	104	37	32	45	84	39	55
	C	41	125	45	92	70	111	109	36	31	98
	D	101	75	43	66	124	73	74	63	72	38
	E	101	51	101	61	70	73	118	62	127	49
	F	88	54	59	52	60	101	128	67	84	54
	G	100	112	40	110	96	106	100	74	111	104
	H	126	122	86	103	80	66	58	35	61	47
	I	30	112	63	115	73	124	125	111	44	66
	J	88	70	58	117	99	80	49	87	97	62

【0 1 5 5】

第2の共起行列についても同様に作成し、実施形態2と同様に以後の処理を行う。具体的には、本実施形態の分析方法、分析装置及び分析プログラムは、共起行列作成手段11が第1の共起行列及び第2の共起行列に基づいて共起行列を作成し、トピック抽出手段12がトピックを抽出し、スコア計算手段13がスコアを計算し、集計手段14が集計を行う。

【符号の説明】

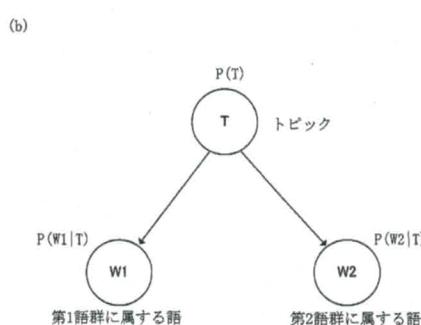
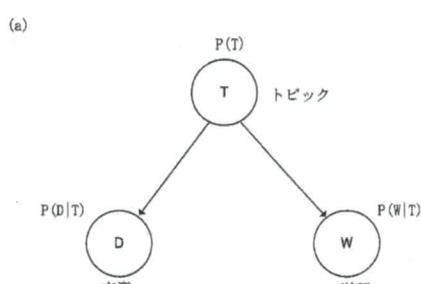
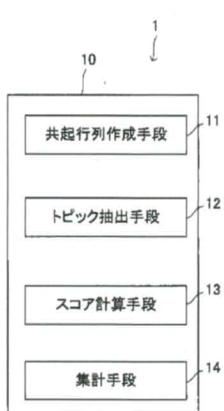
【0 1 5 6】

- 1 分析装置
- 1 0 分析プログラム
- 1 1 共起行列作成手段
- 1 2 トピック抽出手段
- 1 3 スコア計算手段
- 1 4 集計手段

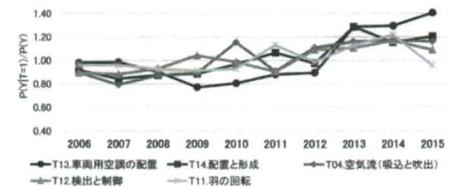
20

【図 1】

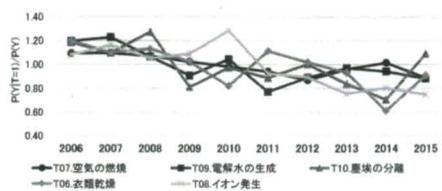
【図 2】



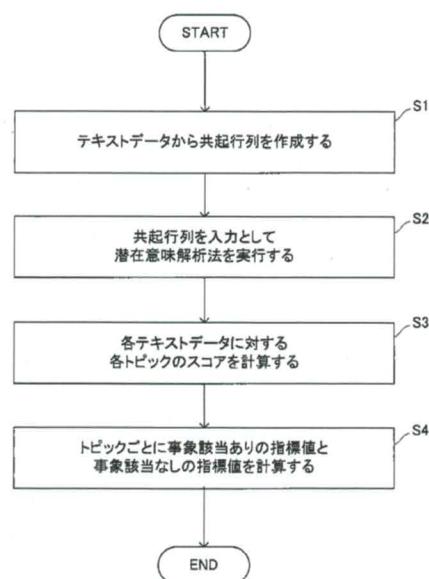
【図 3】



【図 4】



【図 5】



フロントページの続き

(56)参考文献 特開2004-185135 (JP, A)
特開2006-277767 (JP, A)

(58)調査した分野(Int.Cl., DB名)

G 06 F 16/00

G 06 F 40/00