(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.Cl.		F I	
G06F 16/30	*(2019.01)*	G06F	16/30
G06F 16/383	*(2019.01)*	G06F	16/383
G06F 40/20	*(2020.01)*	G06F	40/20
G06F 40/216	*(2020.01)*	G06F	40/216
G06F 40/279	*(2020.01)*	G06F	40/279

請求項の数 10　(全 28 頁)

(21)出願番号　特願2019-84331(P2019-84331)
(22)出願日　平成31年4月25日(2019.4.25)
(65)公開番号　特開2019-200784(P2019-200784A)
(43)公開日　令和1年11月21日(2019.11.21)
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)

(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値を乗じ、
前記補正値は、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、

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値を乗じ、
前記補正値は、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値を乗じ、
前記補正値は、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。
【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値を乗じ、
前記補正値は、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、
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】

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。
【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。
【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。
【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

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値を乗じ、前記補正値は、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。

【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値を乗じ、前記補正値は、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。

【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値を乗じ、前記補正値は、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値を乗じ、前記補正値は、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。

【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】

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

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ＣＰＵ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℃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気浄化装置。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。

【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。

【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。

【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。

【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。

【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。

【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。

【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

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。

【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。

【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】

	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る。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。
【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ましい。
【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）の比率を補正値とする。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。
【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。このような補正は、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値を乗じてもよい。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）の比率を補正値とす

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値を乗じてもよい。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）の比率を補正値とし、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）の比率を補正値とする。

【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。

【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。

【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。

【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（Ｗ１｜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（Ｗ２｜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。

【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。

【００５７】

【表６】

トピックT10			
確率	単語	確率	係り受け
3.4%	塵埃	0.9%	付着-塵埃
1.8%	送風機	0.8%	塵埃-除去
1.7%	掃除機	0.8%	吸い込む-塵埃
1.6%	吸い込む	0.7%	塵埃-含む
1.3%	分離	0.7%	塵埃-吸い込む
1.3%	フィルタ	0.7%	塵埃-分離
1.3%	捕集	0.6%	発生-送風機
1.0%	集塵部	0.6%	含む-空気
…	・・・	…	・・・

トピックT13			
確率	単語	確率	係り受け
1.5%	配置	0.5%	流れる-空気
1.1%	流れる	0.4%	空気-流す
1.0%	向ける	0.4%	備える-車両用空調装置
1.0%	下流	0.3%	空気-送風
1.0%	車両用空調装置	0.3%	下流-配置
0.9%	方向	0.3%	通過-空気
0.9%	車室内	0.3%	前方-配置
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】

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】

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

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Ｔｋを条件とした各文章Ｓｈの条件付確率をＰ（Ｓｈ｜Ｔｋ）とする。各文章の発生確率をＰ（Ｓｈ）とする。各文章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、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】

$$(1)\quad P\left(Sw_h\middle|T_k\right)=\sum_i P\left(Sw_h\middle|W_i\right)P(W_i|T_k)$$

$$(2)\quad P\left(Se_h\middle|T_k\right)=\sum_j P\left(Se_h\middle|E_j\right)P(E_j|T_k)$$

$$(3)\quad P\left(S_h|T_k\right)=P\left(S_h\middle|Sw_h\right)P\left(Sw_h|T_k\right)+P\left(S_h\middle|Se_h\right)P\left(Se_h|T_k\right)$$

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

【００６５】

各文章Ｓｈ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Ｓｗｈ、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Ｓｅｈとする。Ｐ（Ｓｈ｜Ｔｋ）を計算するにあたり、Ｐ（Ｓｗｈ｜Ｔｋ）とＰ（Ｓｅｈ｜Ｔｋ）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Ｗｉが含まれる文章の数をｎ（Ｗｉ）、係り受け表現Ｅｊが含まれる文章の数をｎ（Ｅｊ）とすると、Ｐ（Ｓｗｈ｜Ｗｉ）はｎ（Ｗｉ）の逆数、Ｐ（Ｓｅｈ｜Ｅｊ）はｎ（Ｅｊ）の逆数として計算される。Ｐ（Ｗｉ｜Ｔｋ）とＰ（Ｅｊ｜Ｔｋ）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】

Ｐ（Ｓｈ｜Ｔｋ）は、上記式（３）より得られる。Ｐ（Ｓｈ｜Ｓｗｈ）とＰ（Ｓｈ｜Ｓｅｈ）は文章Ｓｈ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（Ｓｈ）は、上記式（４）で計算され、Ｐ（Ｔｋ）はＰＬＳＡの実行により得られる。

【００６７】

上記式（３）のＰ（Ｓｈ｜Ｔｋ）を、上記式（４）のＰ（Ｓｈ）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（Ｓｈ）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】

このように、Ｐ（Ｓｈ｜Ｔｋ）とＰ（Ｓｈ）との比をもって文章ＳｈにおけるトピックＴｋのスコアとする。この値が１を超えるということは、文章Ｓｈの発生確率はトピックＴｋを条件とすることで上昇し、トピックＴｋ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ＳｈとトピックＴｋの関係の強さを把握しやすくすることができる。表８に各文章Ｓｈに対する各トピックＴｋ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】

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。

【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。

【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（Ｓｈ｜Ｔｋ）／Ｐ（Ｓｈ）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。
【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。
【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】
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。
【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

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ることで得られる。

【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】

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】

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。

【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することができる。

【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値を乗じて補正した。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】

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〉

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。

【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	空気浄化＆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（紙葉類等）	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	洗浄（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。

【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(紙葉類等)
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】

〈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。

【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

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－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。

【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）の比率を補正値とする。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、補正値Ｎ２／Ｎ１を乗じる。

【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。このような補正は、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。

【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。

【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。

【０１２４】

【数２】

$$(1)\quad P\left(Sw_h\middle|T_k\right)=\sum_i P\left(Sw_h\middle|W_i\right)P(W_i|T_k)$$

$$(2)\quad P\left(Se_h\middle|T_k\right)=\sum_j P\left(Se_h\middle|E_j\right)P(E_j|T_k)$$

$$(3)\quad P\left(S_h|T_k\right)=P\left(S_h\middle|Sw_h\right)P\left(Sw_h|T_k\right)+P\left(S_h\middle|Se_h\right)P\left(Se_h|T_k\right)$$

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

【０１２５】

各テキストデータＳｈ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Ｓｗｈ、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Ｓｅｈとする。Ｐ（Ｓｈ｜Ｔｋ）を計算するにあたり、Ｐ（Ｓｗｈ｜Ｔｋ）とＰ（Ｓｅｈ｜Ｔｋ）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Ｗｉが含まれるテキストデータの数をｎ（Ｗｉ）、係り受け表現Ｅｊが含まれるテキストデータの数をｎ（Ｅｊ）とすると、Ｐ（Ｓｗｈ｜Ｗｉ）はｎ（Ｗｉ）の逆数、Ｐ（Ｓｅｈ｜Ｅｊ）はｎ（Ｅｊ）の逆数として計算される。Ｐ（Ｗｉ｜Ｔｋ）とＰ（Ｅｊ｜Ｔｋ）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。

【０１２６】

Ｐ（Ｓｈ｜Ｔｋ）は、上記式（３）より得られる。Ｐ（Ｓｈ｜Ｓｗｈ）とＰ（Ｓｈ｜Ｓｅｈ）は文章Ｓｈ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（Ｓｈ）は、上記式（４）で計算され、Ｐ（Ｔｋ）はＰＬＳＡの実行により得られる。

【０１２７】

上記式（３）のＰ（Ｓｈ｜Ｔｋ）を、上記式（４）のＰ（Ｓｈ）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（Ｓｈ）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】

このように、Ｐ（Ｓｈ｜Ｔｋ）とＰ（Ｓｈ）との比をもってテキストデータＳｈにおけるトピックＴｋ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Ｓｈの発生確率はトピックＴｋを条件とすることで上昇し、トピックＴｋ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ＳｈとトピックＴｋ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Ｓｈに対する各トピックＴｋのスコアを例示する。

【０１２９】

【表１４】

	トピックTk			
テキストデータID	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

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】

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ので、ここでの説明は省略する。

【０１３５】

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】

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℃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

、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。

【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。

【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。

【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

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。

【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ライズされたテキスト部として用いてもよい。

【０１５１】

【表１８】

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℃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】

表１８は、メタデータをカテゴ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】

【表１９】

		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

【０１５５】

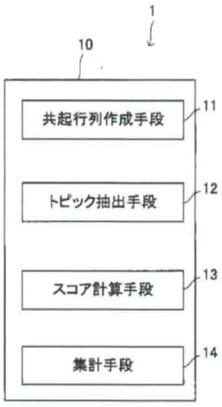
第２の共起行列についても同様に作成し、実施形態２と同様に以後の処理を行う。具体的には、本実施形態の分析方法、分析装置及び分析プログラムは、共起行列作成手段１１が第１の共起行列及び第２の共起行列に基づいて共起行列を作成し、トピック抽出手段１２がトピックを抽出し、スコア計算手段１３がスコアを計算し、集計手段１４が集計を行う。

【符号の説明】

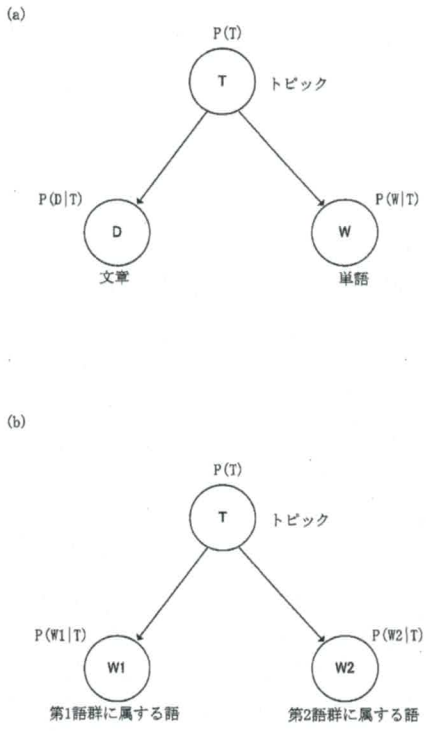
【０１５６】

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
１４　集計手段

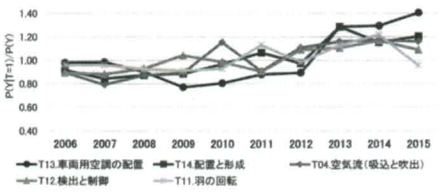
【図１】



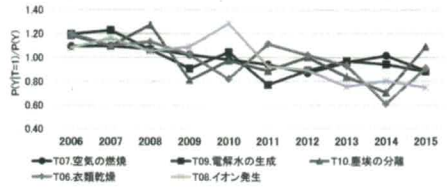
【図２】



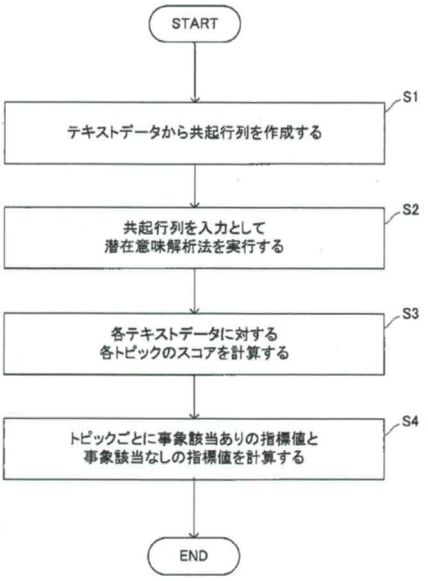
【図３】


1.40
1.20
1.00
0.80
0.60
0.40
P(Y|T=1)/P(Y)
2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
T13.車両用空調の配置
T14.配置と形成
T04.空気流（吸込と吹出）
T12.検出と制御
T11.羽の回転

【図４】


1.40
1.20
1.00
0.80
0.60
0.40
P(Y|T=1)/P(Y)
2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
T07.空気の燃焼
T09.電解水の生成
T10.塵埃の分離
T06.衣類乾燥
T08.イオン発生

【図５】


START
テキストデータから共起行列を作成する
S1
共起行列を入力として
潜在意味解析法を実行する
S2
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
S3
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
S4
END

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)
Ｇ０６Ｆ　１６／００
Ｇ０６Ｆ　４０／００