The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

確率的因果意味解析（PCSA）
－テキストデータを用いたターゲット事象の要因トピックの抽出－

PCSA: Probabilistic Causal Semantic Analysis
Extracting Effective Topics of a Target Event Using Text Data

野守　耕爾[*1]
Koji Nomori

[*1] 株式会社アナリティクスデザインラボ
Analytics Design Lab Inc.

This study proposes a new method extracting topics from text data named PCSA (Probabilistic Causal Semantic Analysis). It enables to extract effective topics of a target event. This paper shows the effectiveness by applying the method to patent data.

1. はじめに

ビジネスにおけるアクションを検討するうえでターゲットとなる事象の要因を探ることは重要であり，例えば商品の顧客満足度の要因を探ることで新商品の企画や現商品のプロモーションを検討したり，店舗来店の要因を探ることで来店を促進する DM を検討したり，サービスの解約や会員退会の要因を探ることで顧客を維持する対応やサービスを検討することができる．

高次元なビッグデータから何かある事象の要因を探る場合，教師なし学習と教師あり学習を組み合わせる手段が有効となることがある．つまり多量な変数を類似したいくつかの特徴に教師なし学習で次元圧縮（グルーピング）し，それを説明変数として目的変数との関係を教師あり学習でモデル化するというアプローチである．従来よく用いられてきたものとして，主成分分析や因子分析をしてその因子を説明変数に回帰分析を実行する手法がある．例えばアンケートの多量な設問項目を因子分析でいくつかの因子にまとめ，顧客満足度などのターゲット設問とその因子との関係を回帰分析して顧客の価値観を理解し，マーケティングの検討に適用されるケースが挙げられる．近年急発展しているディープラーニングも教師なし学習と教師あり学習を組み合わせてモデル化するものである．ディープラーニングでは教師なし学習で特徴量（中間層）を抽出するが，教師あり学習でその特徴量を調整しながら，モデル全体を見直していく手法であり，識別精度を高めることに特化した特徴量が抽出される．

テキストデータの解析でも教師なし学習と教師あり学習を組み合わせることで，テキストデータに潜む要因関係をモデル化する手法が提案されている[野守 2014]．この手法ではテキストマイニングで抽出された単語群に確率的潜在意味解析（PLSA）を適用して教師なし学習によりトピックを抽出し，そのトピックを確率変数としてベイジアンネットワークを適用して教師あり学習によりモデルを構築している．例えば旅行の口コミデータに適用して口コミの評点とトピックとの関係をモデル化したり[野守 2016]，特許の要約文のデータに適用して，技術のトピックと用途のトピックとの関係をモデル化している[野守 2018]．

このように教師なし学習と教師あり学習を組み合わせることで，高次元なビッグデータでも，次元圧縮された特徴量の変数を用いてターゲット事象との関係をシンプルに理解することができる．しかし，このようなアプローチでは教師なし学習と教師あり学習はそれぞれ独立しており，まず教師なし学習を完了してからその結果を教師あり学習に引き継ぐものである．つまり教師なしで抽出された特徴量はどれもターゲット事象と強い関係を示すものではなく，教師あり学習によりその中から関係を示すものが有効な特徴量として選択される．こうした分析をビジネス業務で活用することを考えると，特にターゲット事象に影響を与える要因に集中して抽出することが望ましいとされることが考えられる．ディープラーニングでは先述の通り，教師なし学習と教師あり学習は独立しておらず，教師あり学習が教師なし学習に連携し，ターゲット事象を識別するのに適した特徴量が調整されるが，その特徴量は識別精度を高めることに特化して最適化計算されているため，人間にとって理解しにくいことが多く，しばしばブラックボックスと言われる．

本稿では，テキストデータの活用を想定して，ターゲット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手法を提案し，その適用事例を紹介する．

連絡先：野守耕爾，株式会社アナリティクスデザインラボ，
koji.nomori@analyticsdlab.co.jp

2. 確率的因果意味解析（PCSA）

本研究では，PLSAを応用し，テキストデータからあるターゲット事象に影響を与えるトピックに限定して抽出する手法を PCSA として提案する．（特許出願中，商標出願中）

2.1 確率的潜在意味解析（PLSA）

PLSA(Probabilistic Latent Semantic Analysis)は，文書分類のために開発された次元圧縮手法であり[Hofmann 1999]，トピックモデルと呼ばれる手法の一つである．文書 D とそこに出現する単語 W の間には潜在的な意味クラス C があることを想定し，各文書における単語の出現頻度が記録された「文書」×「単語」の共起行列データを学習し，文書と単語の共通トピックとなるような特徴を見つける手法である．PLSA の実行により 3 種類の確率変数 P(D|C), P(W|C), P(C)が計算され，これにより「文書」×「潜在クラス（トピック）」という低次元データに変換でき，クラスタリングの手法としても用いられる．

PLSA が他のクラスタリング手法と異なる特徴の一つは，行と列を同時にクラスタリングできることである．一般的なクラスタリング手法は，列をベースに行をクラスタリングする，あるいは行をベースに列をクラスタリングするため，どちらか一方しかクラスタリングできない．一方 PLSA で抽出される潜在クラスには，行の要素と列の要素が同時に所属することができる．

2.2 PLSA の共起行列構成の工夫

行と列を同時にクラスタリングできる PLSA では，行と列は双方が十分意味を持つ情報で構成すれば，抽出された潜在クラスの意味を行と列の 2 つの情報軸から解釈することができる．本来の PLSA の適用では，「文書」×「単語」という構成の共起行列をインプットとするが，この構成を工夫することで解釈のしやすい潜在クラス（トピック）を抽出する試みがある．

例えば，「品詞」×「品詞」の共起行列を用いる方法が提案されており，有用な知識が抽出されたことが報告されている[Kameya & Sato 2005][野守 2014]．また，全国の観光地の口コミから得られた「観光地」×「係り受け表現」の共起行列に PLSA を適用することで観光地のテーマを抽出している例もある[野守 2015]．共起行列の軸の一方を「係り受け表現」とすることで，文脈をイメージしやすくトピックの解釈がより容易になったとされており，「単語」×「係り受け表現」の共起行列に PLSA を適用した事例も報告されている[野守 2016] [野守 2018]．これにより単語と係り受けを同時にクラスタリングすることになるが，単語という話題の観点となる軸に基づいて，その観点の具体的な内容となる係り受け表現をグルーピングでき，より文脈上近しい言葉・表現でまとめられた解釈のしやすいトピックを潜在クラスとして抽出できるとされている．

2.3 PLSA の結果を用いた教師あり学習

筆者は，上記のようにテキストデータから PLSA の教師なし学習によって抽出されたトピックに対して，各データの該当有無を計算し，そのトピックを確率変数としてベイジアンネットワークを適用し，教師あり学習によってターゲット事象との関係をモデル化している．例えば温泉旅行の口コミデータから，口コミの満足度（評点）と関係のあるトピックを分析し，観光マーケティングへの応用を検討している[野守 2016]．この適用事例では，まず温泉旅行の口コミ全体で話題にされていることをトピックに集約し，その中から満足度と関係のあるトピックを探索している．しかしトピックは満足度と関係があることを前提に抽出されたものではなく，モデル構築の結果，満足度と関係のないトピックも多く抽出されている．マーケティングなどビジネス業務への応用を想定する場合，特に満足度に影響を与えるトピックに集中して抽出することが望ましいとされるケースが考えられる．

2.4 確率的因果意味解析（PCSA）の提案

本研究では，テキストデータからあるターゲット事象に影響を与えるトピックに限定して PLSA で抽出する手法を，確率的因果意味解析（PCSA: Probabilistic Causal Semantic Analysis）として提案する．

PCSA では，全データから構築した共起行列に対して，あるターゲット事象が起こるデータから構築した共起行列と，それが起こらないデータから構築した共起行列に分割し，その 2 つの共起行列の差分を計算した共起行列に対して PLSA を適用する．PCSA で適用する共起行列のイメージを図 1 に示す．

なお，ターゲット事象が起こるグループと起こらないグループではデータ件数の規模に差が出てしまうため，単純な差分ではなく，件数の少ないグループのデータ件数に合わせて，件数の多いグループの共起行列の各頻度をデータ件数の比率で重み調整する．データ件数の少ないグループに頻度を合わせる理由は，件数の多いグループに頻度を合わせてしまうと，件数の少ないグループの頻度が大きくなるわけだが，特に 1 件などの頻度の少ない共起ペアはたまたま 1 件出現したというケースが大いに考えられ，全体のデータ件数が多くなればその割合に応じて増加すると仮定することは現実と大きく乖離する懸念がある．共起行列では頻度 1 件の共起ペアは最も多く，これがデータ規模に応じて全て同じ割合で増加することの影響は大きいと考えられる．一方，データ件数の多いグループはそのデータ規模でその頻度の共起ペアが存在したことは事実であり，データ件数の少ないグループに合わせて重み調整することでその頻度は小さくなるが，0 件（存在しない）になることはなく，件数の少ないグループの頻度を大きくすることよりも現実的と考えられる．また，2 つの共起行列の差分については，負数とならないようにここでは差の絶対値を取ることにする．

このように設定された共起行列では，ターゲット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ため，この共起行列に PLSA を適用することでターゲット事象の発生有無に影響を与える潜在トピックが優先的に抽出されることが期待される．

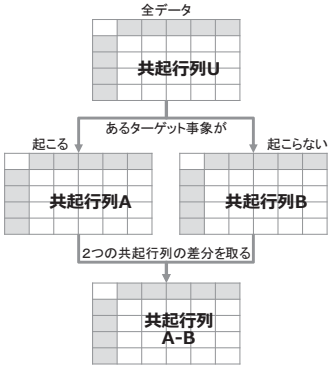

図 1 PCSA で適用する共起行列のイメージ

3. PCSA の適用事例

本稿では PCSA の適用事例として，特許の要約文のテキストデータに適用し，出願が近年上昇傾向，あるいは下降傾向にある技術トピックを抽出した分析結果を紹介する．

3.1 分析データ

特許の要約文と請求項に「風」「空気」を含む 10 年分（出願日が 2006 年 1 月 1 日～2015 年 12 月 31 日）の特許公報 30,039 件を用意し，特にその要約文で「解決手段」に関する項目で記述された文章のデータを分析対象とした．

3.2 PLSA を適用した結果

筆者は本データに対し，「単語」×「係り受け表現」という構成の共起行列を構築し，それに PLSA を適用することで潜在的な技術トピックを抽出し，出願年を軸としたそのトピックのトレンドを分析している[野守 2018]．本節ではその結果の概要を紹介し，次節で PCSA を適用した結果を紹介し，それらを比較する．

3.2.1 抽出されたトピック

まず分析データにテキストマイニングを適用し，単語とその文法的なペアとなる係り受けを抽出した．抽出結果のうち頻度 10 件以上を対象とし，単語（5,187 語）×係り受け（7,174 表現）の共起行列を作成した．各共起ペアの出現を集計する単位は特許単位ではなく文章単位（句点で区切られた文章単位）とした．

この共起行列に PLSA を適用して技術に関するトピックを抽出した．なお PLSA は予めトピック数を設定する必要があり，また初期値により解が異なる特性がある．そこでトピック数を 1 刻みで変化させ，それぞれのトピック数に対して PLSA を初期値を変えて 5 回ずつ実行し，それぞれの解を情報量基準 AIC で評価して最も評価の良い解を採用した．その結果，最も評価の良い解として 47 個のトピックが得られた．

PLSA のアウトプットは，①トピック T における行要素（単語 W）の所属確率 P(W|T)，②トピック T における列要素（係り受け E）の所属確率 P(E|T)，③トピック T の存在確率 P(T)，という 3 つの確率が計算される．特に各トピックにおいて①②の確率の高い単語，係り受けを確認し，そのトピックの意味を解釈していく．抽出されたトピックの内容例を表 1 に示す．トピック t25（表 1 左）では，単語は，水，供給，加湿装置，加湿などが，係り受けは，水の供給，供給される水，空気の加湿などが関係しているので，この結果は加湿に関するトピックであると解釈できる．トピック t32（表 1 右）では，単語は，送風機，塵埃，掃除機，分離，吸い込む，集塵部などが，係り受けは，塵埃の分離，分離する塵埃，塵埃を含む，吸い込む塵埃などが関係しているので，この結果は塵埃の分離に関するトピックであると解釈できる．このように解釈をつけた 47 個のトピックの一覧を表 2 に示す．

表 1　PLSA で抽出されたトピックの内容例

トピックt25			
確率	単語	確率	係り受け
8.5%	水	2.4%	水-供給
2.6%	供給	1.3%	供給-水
2.6%	加湿装置	1.3%	空気-加湿
2.3%	加湿	0.9%	吹出口-連通
1.8%	加湿フィルタ	0.9%	空気-通過
1.7%	水槽	0.9%	連通-空気風路
…	…	…	…

トピックt32			
確率	単語	確率	係り受け
5.5%	送風機	2.1%	塵埃-分離
5.2%	塵埃	1.7%	分離-塵埃
4.1%	掃除機	1.7%	塵埃-含む
3.6%	分離	1.5%	吸い込む-塵埃
3.5%	吸い込む	1.3%	含む-空気
2.3%	集塵部	1.0%	空気-分離
…	…	…	…

表 2　PLSA で抽出された 47 個のトピック一覧

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		

3.2.2 トピックのスコア計算

全特許データに対して各トピックのスコア（該当度）を計算した．1 件の特許は複数の文章で構成されているため，まず文章単位に各トピックのスコアを計算し，それを特許単位に集約した．文章 S_h におけるトピック T_k のスコアは $P(S_h|T_k)/P(S_h)$ で定義した．これはトピックを条件とすることで文章の発生確率が何倍になるのかを示し，そのトピックをよく話題にしている文章ほど高くなる．以下，$P(S_h|T_k)$ と $P(S_h)$ の計算について説明する．

$P(S_h|T_k)$については，文章 S_h を単語で定義される文章 Sw_h と係り受けで定義される文章 Se_h に分解し，それぞれについて $P(Sw_h|T_k)$と $P(Se_h|T_k)$を計算し，それらを一つに統合して $P(S_h|T_k)$を計算する．$P(Sw_h|T_k)$と $P(Se_h|T_k)$は式(1)(2)で計算される．単語 W_i と係り受け E_j が含まれる文章の数をそれぞれ $n(W_i)$ と $n(E_j)$とすると，$P(Sw_h|W_i)$は $n(W_i)$の逆数，$P(Se_h|E_j)$は $n(E_j)$の逆数として計算される．$P(W_i|T_k)$と $P(E_j|T_k)$はそれぞれ PLSA の実行結果によって得られている．$P(S_h|T_k)$は式(3)で計算され，$P(S_h|Sw_h)$と $P(S_h|Se_h)$は文章 S_h において重みは同じであるためそれぞれ 0.5 とする．$P(S_h)$は式(4)で計算され，$P(T_k)$は PLSA の実行によって得られている．

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

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

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

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

以上から$P(S_h|T_k)/P(S_h)$で定義されるスコアを文章単位に計算し，それを特許単位に見たとき，各トピックのスコアの最大値をその特許のトピックスコアとして採用した．さらにこのスコアの閾値として 3 を設定し，各特許データに対してそのトピックの該当有無を示す 0,1 のフラグ情報を付与した．$P(S_h|T_k)/P(S_h)$で定義したスコアは 1 を基準と考えることができるが，各トピックの特徴をより濃く抽出するため，このスコアの分布や実際の文章内容も確認しながら妥当性を検討し，基準の 3 倍と厳しく設定した．

3.2.3 トピックのトレンド分析

特許データの出願年の情報と，各トピックのフラグ情報から，トピックのトレンドを分析した．具体的には出願年 Y とトピック T の関連度を示す指標値として P(Y|T=1)/P(Y)を計算し，この値の出願年変化を集計した．本節では 2013 年を境に，2012 年までのデータ（22,387 件）で計算される指標値と 2013 年からのデータ（7,652 件）で計算される指標値の増減率を計算した結果を紹介する．全 47 個のトピックの増減率の一覧について，増加率の高いものから順に表 3 に示す．表 3 より抽出されたトピックは増減率の高いものから低いものまでばらついていることがわかる．

表 3　47 個のトピックの 2013 年前後における指標値の増減率

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		

3.3 PCSA を適用した結果

本節では，3.2 で紹介した PLSA の適用事例と同一のデータで PCSA を適用した分析結果を示す．ここでの PCSA におけるターゲット事象とは，2013 年以降の特許であるかどうかとし，2013 年前後で上昇傾向にある技術，あるいは下降傾向にある技術に集中してトピックを抽出する．

3.3.1 抽出されたトピック

3.2 の分析事例と同様に，共起行列の構成は単語（5,187 語）×係り受け（7,174 表現）とし，2012 年までのデータ（22,387 件）と2013 年からのデータ（7,652 件）でそれぞれ共起行列を作成した．共起行列の各頻度は文章単位で集計しているが，その文章数は 2012 年までのデータで 33,283 件，2013 年からのデータで 11,831 件あり，2012 年までのデータの共起行列の各頻度を文章数の比率で重み調整した（全ての頻度に 11,831/33,283 を乗した）．この 2 つの共起行列の差の絶対値を計算した共起行列を作成し，これに PLSA を適用した．

3.2 の分析事例と同様に，トピック数を 1 刻みで変化させ，それぞれのトピック数に対して PLSA を初期値を変えて 5 回ずつ実行し，それぞれの解を情報量基準 AIC で評価して最も評価の良い解を採用した結果，14 個のトピックが得られた．抽出されたトピックの内容例を表 4 に示す．所属確率の高い単語と係り受けから，トピック T10（表 4 左）は塵埃の分離に関するトピック，トピック T13（表 4 右）は車両用空調の配置に関するトピックであると解釈できる．トピック T10 は 3.2 の分析事例で抽出されたトピック t32 におおよそ対応するものと思われる．このように解釈をつけた 14 個のトピックの一覧を表 5 に示す．

表 4　PCSA で抽出されたトピックの内容の例

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
…	…	…	…

表 5　PCSA で抽出された 14 個のトピック一覧

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

3.3.2 トピックのトレンド分析

3.2 の分析事例と同様に，全特許データに対して，抽出された各トピックのスコア（該当度）を計算し，スコアの閾値を 3 に設定して，トピックの該当有無を示す 0,1 のフラグ情報を付与した．そして出願年 Y とトピック T の関連度を示す指標値として P(Y|T=1)/P(Y)を採用し，出願年を 2012 年までとした指標値と，2013 年以降とした指標値を計算し，2013 年前後におけるこの値の増減率を集計した．全 14 個のトピックの増減率の一覧について，増加率の高いものから順に表 6 に示す．表 6 を表 3 と比較すると，抽出されたトピックはその多くが増減率の高いものと低いものに集中していることが分かる．また，上記指標値を各出願年において計算し，表 6 で増減率が上位 5 位のトピックと下位 5 位のトピックについて，その経年変化を可視化したものをそれぞれ図 2,3 に示す．今回は 2013 年前後の出願件数に影響を与えるトピックを抽出したが，図 2,3 より全体的に上昇トレンド，下降トレンドにあるトピックが抽出されていることが分かる．

表 6　14 個のトピックの 2013 年前後における指標値の増減率

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

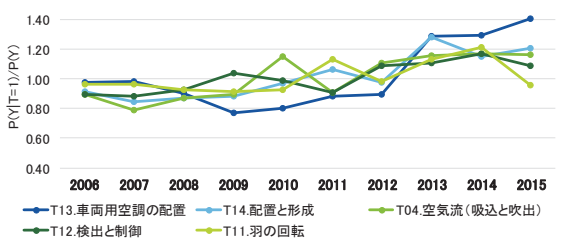

図 2　2013 年前後で増減率上位 5 位のトピックのトレンド

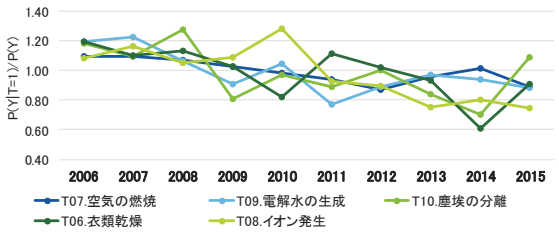

図 3　2013 年前後で増減率下位 5 位のトピックのトレンド

4.　まとめ

本研究では，テキストデータからあるターゲット事象に影響を与えるトピックに限定して PLSA で抽出する手法を，確率的因果意味解析（PCSA）として提案し，同じデータで同じ共起行列の条件のもと，これまでの PLSA を適用した結果と PCSA を適用した結果を示し比較した．その結果，PCSA を適用すると，特にターゲット事象（2013 年前後における特許の出願傾向）に影響を与えるトピックが集中的に抽出されていた．

本手法は様々なテキストデータに適用可能であり，例えば，口コミデータに適用して口コミ評点に影響するトピックを抽出することで満足度を高める商品・サービスを検討たり，コールセンターの問い合わせ履歴に適用し，そのデータに苦情の程度や解約有無などの情報があれば，その要因となる問い合わせトピックを抽出することで，サービスの改善を検討したり，営業マンの日報データに適用して成約に影響する活動のトピックを抽出することで，営業マンの教育に活用することもできる．また，今回の特許データにおいても，例えば調査会社が提供している特許の注目度を示す指標に影響する出願内容のトピックを抽出し，良い特許の要因を探ることも適用アイデアとして考えられる．

このように PCSA はテキストデータに潜む要因関係を顕在化することができる．特にターゲット事象に影響を与える要因を集中的に抽出することができ，効果的なビジネスアクションの検討に有用な知識を提供することが期待できる．

参考文献

[野守　2014] 野守耕爾，神津友武: 三位一体アプローチによるテキストデータモデリング法の開発―宿泊施設の口コミデータを用いた評価推論モデルの構築―，2014 年度人工知能学会全国大会論文集，2014.

[野守　2015] 野守耕爾，神津友武: 口コミデータに PLSA を適用した観光客目線による観光地分析，2015 年度人工知能学会全国大会論文集，2015.

[野守　2016] 野守耕爾，神津友武:口コミビッグデータに人工知能を応用した地域観光の次世代マーケティング―観光客の声に基づいた温泉地の特徴と観光客の価値観の確率モデリング―，2016 年度人工知能学会全国大会論文集，2016.

[野守　2018] 野守耕爾: 人工知能技術を応用した特許文書分析が生み出す新たな技術戦略の検討，経営情報学会 2018 年春季全国研究発表大会要旨集，2018.

[Hofman　1999] Hofmann, T.: Probabilistic latent semantic analysis, Proc. of Uncertainty in Artificial Intelligence, pp. 289-296, 1999.

[Kameya & Sato　2005] Kameya, Y., & Sato, T.: Computation of probabilistic relationship between concepts and their attributes using a statistical analysis of Japanese corpora, Proceedings of Symposium on Large-scale Knowledge Resources, 65-68, 2005.